Computational Modeling and Design of Protein–Protein Interactions

by

Jeliazko R. Jeliazkov

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

June, 2019

Abstract

Protein–protein interactions dictate biological functions, including ones essential to living organisms such as immune response or transcriptional regulation. To fundamentally understand these biological processes, we must understand the underlying interactions at the atomic scale. However interactions are overly abundant and traditional structure determination methods cannot manage a comprehensive study. Alternatively, computational methods can provide structural models with high-throughput overcoming the challenge provided by the sheer breadth of interactions, albeit at the cost of accuracy. Thus, it is necessary to improve modeling techniques if these approaches will be used to rigorously study protein–protein interactions.

In this dissertation, I describe my advances to protein–protein interaction modeling (docking) methods in Rosetta. My advances are based on challenges encountered in a blind docking competition, including: modeling camelid antibodies, modeling flexible protein regions, and modeling solvated interfaces. First, I detail improvements to RosettaAntibody and Rosetta SnugDock, including making the underlying code more robust and easy to use, enabling new loop modeling methods, developing an automatically updating database, and implementing scientific benchmarks. These improvements permitted me to conduct the largest-to-date study of antibody CDR-H3 loop flexibility, which showed that traditional, small-scale studies missed emergent properties.

Then, I pivot from antibodies to focus on the modeling of disordered protein regions. I contributed advances to the FloppyTail protocol, including enabling the modeling of multiple disordered regions within a single protein and pioneering an ensemble-based

analysis of resultant models. I modeled Hfq proteins across six species of bacteria and demonstrated experimentally-validated prediction of interactions between disordered and ordered protein regions. My simulations provided a hypothetical mechanism for Hfq function.

Finally, I designed crystallographic protein–protein interactions, with the goal of improving protein crystal resolution. To approach this exceptional challenge, I first demonstrated that, under homogenous conditions, Rosetta scores can correlate with crystal resolution. Next, I computationally designed and experimentally characterized sixteen variants of a model protein. Only five crystallized, with one providing an improvement in resolution, showing that improvement through computational design is challenging, but possible.

In sum, my work advanced our understanding and our ability to model and design several challenging protein–protein interactions.

Thesis Committee

Jeffrey J. Gray (Primary Advisor)
Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins Whiting School of Engineering

James M. Berger (Reader)
Professor
Department of Biophysics and Biophysical Chemistry
Johns Hopkins School of Medicine

Bertrand García-Moreno E.
Professor and Chair
T.C. Jenkins Department of Biophysics
Johns Hopkins Krieger School of Arts and Sciences

Margaret E. Johnson
Assistant Professor
T.C. Jenkins Department of Biophysics
Johns Hopkins Krieger School of Arts and Sciences

Jamie B. Spangler
Assistant Professor
Department of Biomedical Engineering
Johns Hopkins Whiting School of Engineering

Acknowledgments

I would like to begin by thanking *you all*: every person I have had the pleasure of knowing throughout my life. This way no one who has directly or indirectly (but still significantly) contributed to this dissertation can say I have forgotten them. With that safety net in place, I will give some more specific thanks.

The easiest thanks to give is to my advisor, Professor Jeffrey J. Gray. Were I to list all of Jeff's positive attributes and how they contribute to his excellence as mentor, I would run out of paper. I think my favorites are Jeff's infectious optimism and scientific curiosity. Through these Jeff cultivates an unbelievable feeling of worth and purpose in his students and their scientific research. As student, he makes it easy to get excited about your work, turning the long slog that is graduate school into a rather fun time!

He also cultivates a lab environment that is ineffable, incomparable, indescribable, irreplaceable, and I think I'm out of "i" words, but trust me it is something special. A large part of that is due to my labmates, past and present. Thanks folks. My first contact within the Gray lab was Daisuke Kuroda, who rather begrudgingly decided to take a gamble on me as a rotation student. In the end, it was less of a gamble and more of a long term investment, as we eventually (four years later) published results from the rotation project he proposed. Along the way Daisuke set an exceptional example for how to perform rigorous, comprehensive, and reproducible research. When I joined the Gray lab, I met the other members (at that time): Brian Weitzner, Sergey Lyskov, Julia Koehler Leman, Jason Labonte, Shourya Sonkar Roy Burman, Michael Pacella, Krishna Kilambi, and Nicholas Marze.

I do not think a sentence or two can do these wonderful people justice but I will try. Brian is the only person in the world who can make reading manuals cool. He constantly strives for perfection in everything and inspires me to do the same, whether it is developing code, writing a paper, or just living your life. I am lucky that he, to this day, remains responsive to my solicitations for advice, even though he graduated years ago. Likewise, Julia is a current colleague and collaborator, without her I would not have published nearly as many papers. She is fearless in her pursuit of science and was largely responsible for showing me that the Rosetta community, and that developing code in a team, was not a scary endeavor. With Jason and Mike, I was unfortunate to not have much scientific overlap, as it would have been quite fun to work with them professionally. Instead, we chose to collaborate on non-professional outdoors experiences, including my first ever camping trip (featuring freezing temperatures and snow in the Pemigewasset Wilderness in early March). Shourya on the other hand was unfortunate to have scientific overlap with me. We schemed away at several challenging modeling problems together. In the end, we were a quite good team: his rational and fundamental approach to modeling complemented my wild and creative one. Finally, Nick and Krishna were my first office mates. Krishna graduated soon after I joined the lab, but he imparted a permanent impression on the lab in the form of "Krishna time"[1]. Nick and I overlapped for about four years, during which we spent many "work" hours debating the solution to the weekly FiveThirtyEight Riddler. But to paint a picture solely of indolence would be inaccurate. Nick is one of the smartest individuals I know (in fact I expect him to win Jeopardy! one day) and I have been lucky to have had access to him for years. Our interactions have shaped my approach to asking scientific questions for the better.

As the years have gone by, the composure of the Gray lab has changed and I have been happy to meet many new colleagues, including Elizabeth Lagesse, Rebecca Alford,

[1]Krishna time encompassed Krishna's atypical work hours and his steadfast belief that the time set aside for a meeting included the time required to commute to the meeting. This resulted in a bit of a surprise for me when I attended my first Gray lab meeting on time and everyone else arrived 15 minutes later.

Morgan Nance, Kayvon Tabrizi, Naireeta Biswas, Joseph Lubin, Xiyao Long, Paige Stanley, Aleexsan Adal, Brittany Lasher, Sai Pooja Mahajan, Jing Zhou, Sudhanshu Shanker, and Ameya Harmalkar. Morgan and Elizabeth provided some biophysics backup in a lab dominated by chemical engineers. Pooja and Ameya have proven themselves as apt office-mate replacements for Nick and Krishna, although our debates are now more work- rather than Riddler-oriented. Finally, among the litany of Gray lab members, I particularly enjoyed mentoring Jing, Paige, and Xiyao, all of who made it easy by being brilliant mentees.

I have also benefited greatly from "science friends" or collaborators that are more than that. The tone was set by Dr. Andrew Santiago-Frangos and Prof. Sarah Woodson (both of JHU). Over four years we collaborated on two publications and I had countless scientific debates about computation with Andrew. These debates substantially improved my ability to communicate my research to non-computational scientists. Another two-paper collaboration involved Rahel Frick and Prof. Inger Sandlie from the University of Oslo. In fact, Rahel enjoyed working with us so much that she has decided to join the Gray lab as a postdoc (to be fair I almost went to Oslo for the same reason)! Finally, Profs. James Berger and Bertrand García-Moreno risked lab resources and lab member effort to enable me to learn crystallography and other experimental techniques. Without them and members of their labs (thanks Matthew Hobson and Aaron Robinson) an entire chapter of this dissertation would not have been possible.

Along those lines, I want to thank my friends outside of science and the lab. I owe a lot to them, especially the ones who have had to put up with me the longest. Growing up, I enjoyed the company of Vincent Bonanno, Danny and TJ Santoro, Tyler Bannan, Mike Thompson, Bill Nguyen, Nicole Schiavone, and Tara Moken. All of who have done well to keep in touch ever since the days of grade school. Throughout college, I was kept honest by Jon Durfee, Natalie Cowan, Ikoro Ikoro (not a typo), and Ricardo Rodriguez. They are also a group that does not lose touch.

And then there is the group that is in constant touch: my roommates. They are epito-

mized by Sean Klein, who has been my roommate for the last six years (poor him). Sean is not only an outstanding human being, but also a very eager conversationalist. Over the years I have enjoyed many discussions, several heated debates, and the occasional banter with Sean. I am afraid I will have to live alone for the rest of my days, since no one else is likely to match how good of a roommate he has been. Our third roommate, one of Lauren Que, Nathan Klein, and Matthew Hobson depending on the year, has not only put up with us brilliantly, but also provided quite good company of their own.

Surprisingly, people also spent time with me by choice rather than because we shared an apartment. Mariusz Matyszweski (whose name I certainly spelled correctly and without looking up), Andrew Santiago-Frangos, Henry Lessen, and Max Klein, have wasted a lot of time on me and been my board game, video game, bar trivia, and drinking buddies for years. Alongside them, I've also had pleasure of knowing and enjoying the company of Chris Borher, Laura Nevin, Ryan McQuillen, Sarah Kim, Cameron Avelis, Dagan[2] Marx, Matthew Sternke, Emily Grasso, Christos Kougentakis, Dillon Nye, Miranda Russo, Meredith Peck, Joseph Rehfus, and Sabrina Schatzman. On a more healthy note, I have had joy of participating in numerous intramural and social soccer leagues with Marco Chiaberge, Kyle Roschli, and Chris Duckworth, in addition to a few of the aforementioned.

Personally, I find it exceptionally dissatisfying to agglomerate the last two paragraphs' worth of individuals into lists of names. I could write paragraphs about each of my graduate school friends before exactly capturing their influence on my life. Honestly, you folks have improved me beyond belief and who I am will always be comprised of a little contribution from you.

Maybe it is Bulgarian culture, or possibly it is luck, but I am extremely thankful to not only have an amazing family but also truly wonderful American-based, Bulgarian family friends. I have actually never written these names in English, so this will be fun: Hristina and Ivan Ivanovi, Petar and Iveta Pirgovi, Sylvia and Momchil Monovi, Joro and Emilia

[2] Also spelled Dagen.

Ekimovi, Severin and Lucy Severovi, Bobby and Tanya Todorovi, and Plamen Todorov. And I should not forget to mention my fellow American-Bulgarian kids: Lyuba Pirgova, Tzvetan and Tanya Monovi, Victoria Todorova and Galen Ekimov. It takes a village to raise a child after all, and my village was the best.

Finally, I would like to thank my family for putting up with me all these twenty-some-odd years. My maternal, Anna and Kostadin, and paternal, Valentina and Zhelyasko, grandparents cared for my sister and me pretty much every summer until we went off to college. They taught us most of what we know about our culture and heritage, including how to read and write in Cyrillic, despite our strong resistance at the time. Only now do we realize how valuable those lessons were and how much they have enriched our lives. They did not do it alone. Our extended family was instrumental in helping them. Aunts and uncles, Petar and Lily and Rumyana and Petar (yes, another Petar); cousins, Simeon, Anne, and Danail; godparents, Anka and Petar (third time is the charm); and family friends, in particular Sasho, all contributed to our cultural upbringing in one way or another.

For all the suffering involved in growing up alongside me, my sister Valentina has earned at the very least her own paragraph. We were a rambunctious pair of children. I vividly recall that we once spent a summer quite literally climbing up the walls. In fact, Valentina was so integral to my childhood that it is difficult to recall a memory without her, and she remains integral to my life today.

I certainly would not be writing this dissertation were it not for my parents, for a number of reasons (including the obvious). They very bravely risked it all to immigrate here when my sister and I were still young. Since then, they have spent every moment working to give us the best opportunities and highest quality of life possible. Through their hard work, they have enabled my pursuit of a PhD. I do not think I know anyone who works harder than they do. In fact, they are not just paradigms of work ethic, but rather they set superlative examples in every aspect of life for us. If I could emulate 5% of how my parents live and what they have done in their lives, I would be happy. *Obicham ve*.

Life is short,

and art long,

opportunity fleeting,

experimentations perilous,

and judgment difficult.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

> Molecules orchestrate the processes of life: complex, diverse molecules with focused functional capabilities. Like the machines of the modern world, these molecules are built to perform specific functions efficiently, accurately, and consistently.
>
> David S. Goodsell, *The Machinery of Life*

1.1 Proteins are integral to biological functions

Essential to life are proteins: unbranched polymers of variable length composed of amino acid building blocks. After water, amino acids are the second largest contributor to cell mass[1]. The sequence of amino acids determines a protein's three-dimensional structure and function. There are twenty "canonical" amino acids, with chemically diverse properties (varying in size, charge, hydrophobicity, and hydrogen bonding character). In water, a protein will "fold" to bury the hydrophobic amino acids and expose the hydrophilic amino acids, and form hydrogen bonding and electrostatic interactions across amino acids. The structure of a "folded" protein will determine its function: a protein with enzymatic activity might have binding pockets for reactants that coax the molecules into a geometry favorable for the forward reaction, a structural protein might expose two patches of complementary charge that lead to the formation of long polymers by end-to-end stacking of multiple molecules, or a protein involved in homeostasis might use a hydrophobic patch to recognize and sequester an unfolded protein.

As these examples imply, proteins do not exist in isolation, and protein–protein interactions are exceptionally prevalent. In humans for example, it is estimated that approximately 250,498 protein–protein interactions exist among 10,531 studied proteins[2]. The actual number of interactions is likely much higher as there are approximately 20,000 protein-coding genes[3] and that study did not consider interaction between human and non-human proteins. In light of their breadth, the study of protein–protein interactions may seem futile. However, interactions can be grouped, categorized, and classified, and studies can prioritize interactions of significant biological importance. In this dissertation, I focus on a subset of interactions that are challenging to study with current methods.

1.1.1 Protein structure is determined through experimental methods

Traditionally, to better understand the role of proteins and their interactions in fundamental biological processes, scientists have turn to experimental structure determination[4–10]. X-ray crystallography is the premier method[i] for acquiring atomic-resolution protein structures. However, it requires an extremely pure sample of protein that has been coaxed to form a crystalline state, such that all molecules of the protein adopt identical conformations in a symmetrically repeating fashion[ii]. Producing such a sample can require significant time and resource investment and does not always guarantee successful structure determination.

In my PhD, I sought to develop computational methods to supplement experimental structure determination methods. For example, the significant cost in time means that protein targets must be specifically selected and cannot be studied in a high-throughput fashion. This shortcoming prevents the effective study of systems where numerous proteins differ slightly in sequence but greatly in structure and function. I developed and applied computational modeling approaches to study two such systems: the millions of unique antibodies in the adaptive immune system and Hfq proteins, which are ubiquitous across

[i] Two alternative approaches are nuclear magnetic resonance (NMR) and (cryo-) electron microscopy (cryo-EM or EM).

[ii] NMR and cryo-EM have their own shortcomings: NMR struggles with large proteins, whereas cryo-EM struggles with small proteins and cannot reliably acquire high-resolution data

many bacterial species. Computational modeling provides an attractive alternative for studying these systems, due to its time and cost efficiency. Additionally, as the quality of the protein crystal is a key factor that determines the resolution of the data, I sought to computationally design high-resolution protein crystals.

1.1.2 Antibodies protect vertebrates from foreign pathogens

Antibodies are proteins produced by the vertebrate adaptive immune system in specific response to pathogenic molecules (antigens). Their function is to target an antigen (by specific binding) for destruction (by activating the complement system or signaling for phagocytosis/degranulation)[11]. Antibody diversity mediates the specific binding of numerous antigens and arises from genetic mechanisms. Briefly, the prototypic[iii] antibody is a pair of pairs of proteins, where a heavy and light chain pair with another, identical heavy and light chain pair (Figure 1.1). The N-terminal domains of the chains are highly variable. In the case of the human heavy chain, one combination is produced from 56 V, 23 D, and 6 J genes, in a process known as VDJ recombination. For the light chain this process is similar, except there is no D gene and there are two sets of V/J genes (referred to as λ and κ, with 205 and 165 combinations, respectively). The heavy VDJ-recombined gene is paired with one of nine constant domain genes (which vary in length), whereas the light VJ-recombined gene is paired with one of five constant domain genes (which do not vary in length). When expressed, heavy and light chains are paired with each other to produce a naïve antibody (one that is not specific to an antigen). Such antibodies are displayed on B cells and when they begin to bind antigen (in the presence of antigen-specific helper T cells), a process known as somatic hypermutation occurs, which introduces point mutations in the variable region. Simultaneously, there is a selection pressure for antibodies that bind antigen specifically and with high affinity.

The diversity at the genetic level gives rise to a structural diversity. A single antibody

[iii] IgG isotype

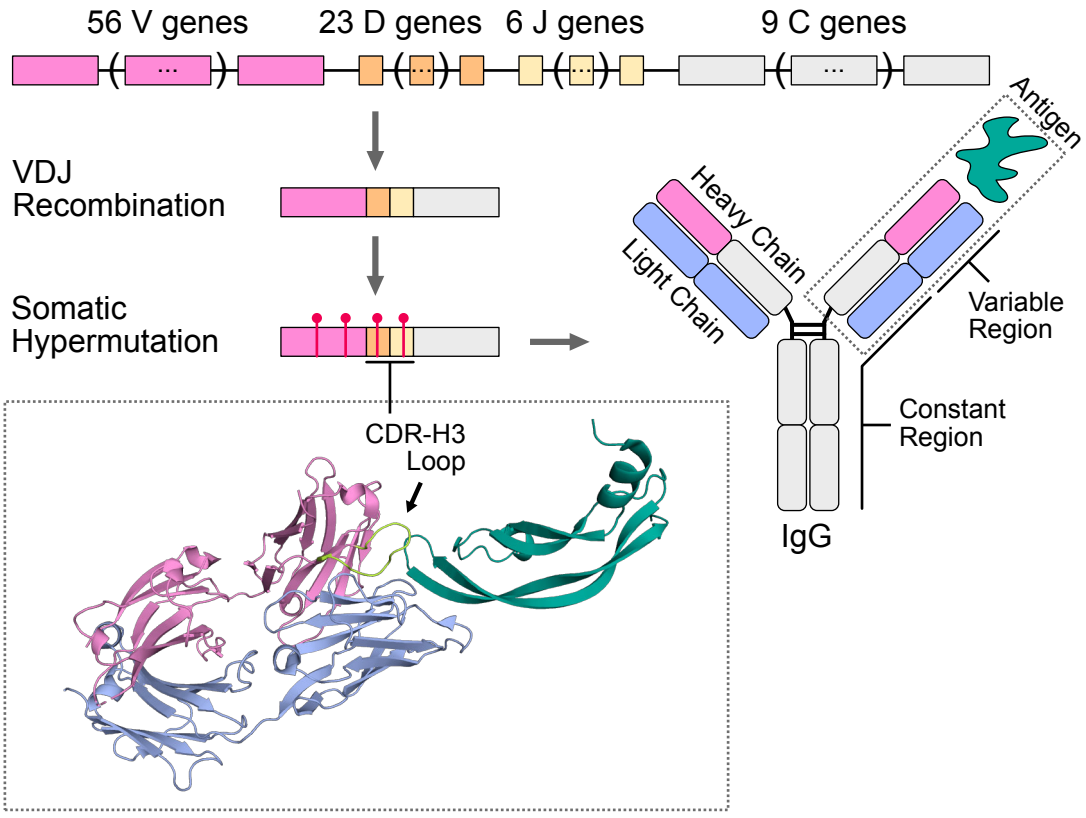

Figure 1.1: Antibody proteins are generated by the recombination of V, D, and J genes for the heavy chain and V and J genes for the light chain (not shown), followed by somatic hypermutation of the genes under a selective pressure for antigen binding. The antibody isotype is determined by the constant domain gene. Gene counts are from IGMT queries[12] for functional *Homo sapiens* genes in IMGT groups: IGHV, IGHD, IGHJ, and IGHC . The dashed box shows a crystal structure (PDB ID 3BDY) of the an antibody Fab fragment (variable region and the first constant domain) bound to an antigen (vascular endothelial growth factor, VEGF).

chain contains multiple immunoglobulin domains. A single immunoglobulin domain is approximately 130 residues in length and contains two anti-parallel β-sheets sandwiched against one another, with a single cross-sheet disulfide bond. The strands that make up the sheets are connected by flexible loops. The subset of these loops that is solvent exposed and oriented away from the first constant regions is the source of structural diversity in the variable region[13]. In particular, there are three key loops, referred to as complementarity determining regions (CDRs) for their role in antigen recognition. In most antibodies, the CDRs are the regions making contacting with the antigen (*i.e.* the CDRs are the paratope)[14].

Not all CDRs contribute equally to antigen interactions. The third loop on the heavy chain often contributes the majority of antigen-binding energy[15]. Unsurprisingly, the CDR-H3 loop is the center of VDJ recombination, containing the entirety of the D gene and parts of the V and J genes. Diversity at the genetic level gives rise to structural plasticity: 30% of CDR-H3 loops have unique structures when compared to a set of non-redundant protein loops whereas that number is only 3% for all other loops[16]. While the variation across antibody structures and sequences is essential to their function, it precludes comprehensive study. Uncovering the relationship between antibody sequence and structure, and how it contributes to antigen binding, would enable more efficient development of vaccines and therapeutics.

1.1.3 Hfq facilitates RNA–RNA interactions

Hfq (Host factor for RNA phage Qβ replication) is an RNA-binding protein present in most sequenced bacteria[17]. Its role is to regulate the expression of metabolic, stress-response, and virulence genes[18] by facilitating interactions between small non-coding RNAs (sRNAs) and their cognate mRNA. Hfq must rapidly anneal RNAs and dissociate the product in an efficient manner, as there are many cellular nucleic acids[19]. The RNA-binding sites in the core, folded portion of Hfq are well studied[20–22], but less is known about the functional importance of its disordered termini.

In the context of proteins, disorder implies the lack of ordered structure (in the absence of external perturbation). Instead, disordered proteins or protein regions occupy an ensemble of conformational states. Despite seemingly breaking the paradigm that structure determines function, disordered proteins and protein regions are functional, and estimates for the disorder content of a given proteome vary from 10–40%[23]. Furthermore, disordered proteins have significant biological relevance as they are the causative agents of many forms of neurodegenerative disease[24].

In Hfq, a recent study revealed that the disordered C-terminal domain (CTD) can displace RNA from the Hfq protein[25], but did not identify the molecular mechanism. As the termini are conserved across bacterial species[26], but vary in their sequence and length (as is common for disordered peptides), it is appealing to suggest that the CTD behaves according a random polymer model and randomly, or non-specifically, displaces Hfq-bound RNA. The alternative hypothesis is that there are specific interactions guiding the CTD to the RNA-binding sites and that RNA displacement is not random. Developing a method to distinguish the two possibilities would be potentially useful for predicting the sequence–function relationship of disordered domains in other partially disordered proteins.

1.1.4 Protein crystals form through repetitive, identical protein–protein contacts

In order to determine a protein structure through X-ray crystallography, one needs a protein crystal. First, a pure sample of the target protein must be produced at a medium-to-high concentration, typically 15–20 mg/ml. Next, one of several approaches could be taken to grow a protein crystal. In the vapor diffusion method, the purified protein is mixed with a precipitant and a small volume (drop) of the mixed solution is placed in a sealed container with a reservoir solution containing a higher precipitant concentration. Through diffusion, the drop and reservoir come into equilibrium. This action increases the precipitant concentration in the drop, forcing the protein to supersaturate. The supersaturated state

is not stable, so the protein that is in excess of its solubility limit solidifies, either by aggregating in an unstructured fashion or nucleating a crystal[27]. Following nucleation, the protein crystal grows until the on-rate for molecules diffusing to and encountering the crystal in the correct orientation matches the off-rate for molecules detaching from the crystal, which is related to the strength of the protein–protein interaction in the crystal lattice and the protein concentration in solution. The crystal is then harvested, frozen, and diffraction data is collected.

The resolution of the structure determined from the diffraction data depends on multiple variables: the flux of the X-ray beam, the detector size and pixel count, and the quality of the crystal. If a protein crystal is conformationally heterogeneous (*i.e.* the repeating units are not truly identical) or if it is anisotropic (*i.e.* grows asymmetrically along an axis), then diffraction data will be of lower quality than for a well-ordered, isotropic crystal. As both of these properties depend on the protein–protein interactions in the lattice[iv], it should be possible to develop computational design strategy to strengthen crystal contacts (when they are known) and improve crystal resolution.

1.2 The Rosetta software suite

Rosetta is a software suite for biomolecular structure prediction and design. Initially, Rosetta was developed to predict protein structure, given protein sequence[29]. Then, it was applied to protein design (*i.e.* the reverse process: given protein structure, predict the optimal sequence)[30]. As Rosetta demonstrated success in both of these domains, it was applied to more challenging and diverse problems, including the prediction[31] and design[32] of protein complexes. Nowadays, the structural modeling or design of exotic molecules such as carbohydrates[33] or RNAs[34] is possible.

At its core Rosetta has two primary functions: (1) sampling relevant conformational

[iv] Conformational heterogeneity can also arise from internal protein flexibility, in which case other engineering approaches such as surface entropy reduction can be taken[28].

space, *e.g.* the native or folded state, and (2) distinguishing this native conformation from the many possible conformations sampled during modeling. Rosetta samples conformational space in a Monte-Carlo-plus-Minimization fashion. The protein, complex, sugar, RNA, *etc.* degrees of freedom are perturbed, the system degrees of freedom are energy-minimized, and changes are accepted or rejected according to the Metropolis criterion: accept if $A \geq U(0,1)$ where $U(0,1)$ is a uniform random number between zero and one, inclusive, and $A = \min(1, e^{-\Delta E/kT})$, with ΔE being the change in energy between the initial and final conformations. The energy of a biomolecular conformation is computed by a hybrid physics- and statistics-based scoring function in Rosetta[35].

1.2.1 Rosetta samples in internal coordinate spate

Rosetta is constructed in a object-oriented fashion. At the heart of Rosetta lie "mover" objects which make conformational changes to "pose" objects (collections of molecules). For the sake of efficiency, Rosetta uses internal coordinates (ϕ, ψ, ω and χ angles for *each residue*) to store protein conformations, instead of traditional Cartesian coordinates (x, y, and z for *each atom*). Thus, the primary degrees of freedom sampled during a simulation are dihedral angles and rigid-body transformations (if multiple proteins are present). The bond lengths and angles could be sampled but are typically held fixed.

1.2.2 Rosetta scores with a hybrid statistical/physical potential

Rosetta combines a milieu of physical and statistical terms to approximate the energy of a biomolecular conformation[35]. The total energy is computed as a sum over the terms, which are a function of the degrees of freedom of the system (D) and the chemical identities (aa): $E_{\text{total}} = \sum w_i E_i(\text{D}, \text{aa})$, where physical terms are weighted at 1.0 and statistical terms are optimized to reproduce small-molecule thermodynamics[36] and features of high-resolution protein crystal structures[37]. The standard Rosetta scoring function (`talaris2014` for studies prior to 2016 and `REF2015` afterward) includes the following:

- a Lennard-Jones potential,
- a Gaussian exclusion implicit solvation potential (with orientation-dependent solvation of polar atoms added in `REF2015`),
- a Coulombic electrostatic potential,
- an orientation-dependent hydrogen bonding potential,
- disulfide-bond potential,
- a statistical potential for the amino acid identity (given the backbone dihedral angles),
- a statistical potential for the backbone dihedral angles (given the amino acid identity),
- a statistical potential for the side-chain rotamer (given the backbone dihedral angles),
- a term penalizing deviation from a planar peptide bond ($\omega = 0$ °or $\omega = 180°$),
- a term penalizing the opening of proline ring: a term penalizing a non-planar tyrosine $\chi 3$ dihedral angle, and
- a set of unfolded-state reference energies for each amino acid identity.

1.2.3 Rosetta modeling is assessed on known structures.

In a standard Rosetta modeling protocol, thousands of models are produced and scored. The set of lowest-scoring models is assumed to be representative of the native structure. This is not speculation. As Rosetta protocols are developed, their performance is validated against proteins with known structures such that the root-mean-squared-deviation (RMSD) between model and structure coordinates can be computed. The RMSD is plotted against the score for the entire set of models (as in Figure 1.2). For successful simulations, low-RMSD (native) models are also low-scoring models whereas high-RMSD (non-native) do not score well. This gives the figure a funnel-like shape and RMSD vs. score plots are typically referred to as funnel plots.

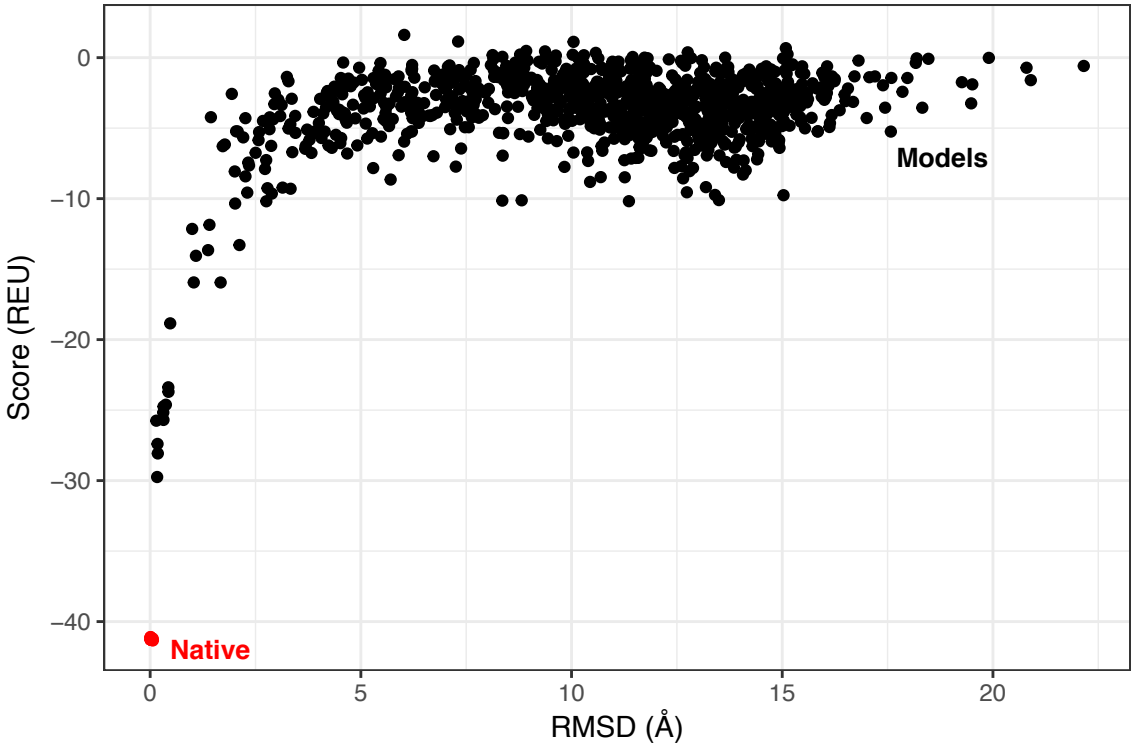

Figure 1.2: A sample funnel plot of the antibody-bound tissue factor extracellular domain (PDB ID 1AHW). The interface energy (y-axis, score of the complex minus score of the individual partners) is plotted against the Cα RMSD (x-axis). Ten "native" models, starting from the crystal structure and refined in the Rosetta score function, are shown for reference in red. One thousand models from a docking simulation are shown in black.

1.3 Dissertation outline

In the present thesis, Chapter 1 provides an introduction to the general themes of my research. Next (Chapter 2), I highlight interesting and challenging targets from my participation in the Critical Assessment of PRotein Interactions (CAPRI) competition over the last six years. A set of challenges was associated with robustness and assumptions underlying the modeling of antibodies and antibody–antigen interactions in Rosetta. In Chapter 3, I describe the technical advances made to simplify the use of RosettaAntibody and Rosetta SnugDock from the user perspective and to expand the classes of antibodies to which these tools can be applied. In Chapter 4, I apply RosettaAntibody and graph theory to assess the effects of affinity maturation on CDR-H3 loop flexibility. The analysis of a large set of antibodies permitted me to observe new, emergent properties, which were not evident in previous, small-scale studies. In Chapter 5, I continue to apply Rosetta to model structural properties at a larger scale than is possible with experiment, this time studying the Hfq family of proteins. I identify key atomic interactions between Hfq's ordered core and disordered termini that span multiple bacterial species and can be tied directly to its function. In Chapter 6, I expand on my structural modeling expertise by attempting to design crystallographic protein–protein interactions. I demonstrate that point mutations can substantially alter protein crystal resolution, although there is not yet a way to reliably predict the exact effect of each mutation. Finally, I map my contributions to the field and detail directions for future explorers in Chapter 7.

References

1. Alberts, B. *et al. Essential Cell Biology* 3rd (Garland Science, New York, 2009).
2. Kotlyar, M. *et al.* In silico prediction of physical protein interactions and characterization of interactome orphans. *Nature Methods* **12,** 79–84 (2015).
3. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
4. Lawson, D. M., Smith, J. M. A. & et al. Solving the Structure of Human H Ferritin by Genetically Engineering Intermolecular Crystal Contacts. en. *Nature* **349,** 541–544 (1991).
5. Chen, P. *et al.* Structure of the human cytomegalovirus protease catalytic domain reveals a novel serine protease fold and catalytic triad. *Cell* **86,** 835–843 (1996).
6. Kim, C.-Y., Quarsten, H., Bergseng, E., Khosla, C. & Sollid, L. M. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 4175–9 (2004).
7. Skordalakes, E. & Berger, J. M. Structural insights into RNA-dependent ring closure and ATPase activation by the Rho termination factor. *Cell* **127,** 553–64 (2006).
8. Zhu, K. *et al.* Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins: Structure, Function, and Bioinformatics* **82,** 1646–1655 (2014).
9. Morag, O., Sgourakis, N. G., Baker, D. & Goldbourt, A. The NMR–Rosetta capsid model of M13 bacteriophage reveals a quadrupled hydrophobic packing epitope. en. *Proceedings of the National Academy of Sciences,* 201415393 (2015).
10. Matyszewski, M., Morrone, S. R. & Sohn, J. Digital signaling network drives the assembly of the AIM2-ASC inflammasome. *Proceedings of the National Academy of Sciences of the United States of America* **115,** E1963–E1972 (2018).
11. Murphy, K., Weaver, C. & Mowat, A. *Janeway's Immunobiology* 9th Editio, 1–907 (2017).
12. Lefranc, M.-P. *et al.* IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research* **27,** 209–212 (1999).
13. Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. & Winter, G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* **321,** 522–525 (1986).
14. Sela-Culang, I., Kunik, V. & Ofran, Y. The structural basis of antibody-antigen recognition. *Frontiers in Immunology* **4,** 1–13 (2013).

15. Kunik, V. & Ofran, Y. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. en. *Protein Engineering, Design and Selection* **26,** 599–609 (2013).

16. Regep, C., Georges, G., Shi, J., Popovic, B. & Deane, C. M. The H3 loop of antibodies shows unique structural characteristics. *Proteins: Structure, Function and Bioinformatics* **85,** 1311–1318 (2017).

17. Sun, X., Zhulin, I. & Wartell, R. M. Predicted structure and phyletic distribution of the RNA-binding protein Hfq. *Nucleic Acids Research* **30,** 3662–3671 (2002).

18. Feliciano, J. R., Grilo, A. M., Guerreiro, S. I., Sousa, S. A. & Leitão, J. H. Hfq: a multifaceted RNA chaperone involved in virulence. *Future Microbiology* **11,** 137–151 (2016).

19. Rajkowitsch, L. *et al.* RNA Chaperones, RNA Annealers and RNA Helicases. *RNA Biology* **4,** 118–130 (2007).

20. Schumacher, M. A., Pearson, R. F., Møller, T., Valentin-Hansen, P. & Brennan, R. G. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: A bacterial Sm-like protein. *EMBO Journal* **21,** 3546–3556 (2002).

21. Zhou, L. *et al.* Crystal structures of the Lsm complex bound to the 3âĂš end sequence of U6 small nuclear RNA. *Nature* **506,** 116–120 (2014).

22. Mikulecky, P. J. *et al.* Escherichia coli Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. *Nature Structural & Molecular Biology* **11,** 1206–1214 (2004).

23. Oates, M. E. *et al.* D2P2: database of disordered protein predictions. *Nucleic Acids Research* **41,** D508–D516 (2012).

24. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. *Annual Review of Biophysics* **37,** 215–246 (2008).

25. Santiago-Frangos, A., Kavita, K., Schu, D. J., Gottesman, S. & Woodson, S. A. C-terminal domain of the RNA chaperone Hfq drives sRNA competition and release of target RNA. *Proceedings of the National Academy of Sciences of the United States of America* **113,** E6089–E6096 (2016).

26. Zheng, A., Panja, S. & Woodson, S. A. Arginine Patch Predicts the RNA Annealing Activity of Hfq from Gram-Negative and Gram-Positive Bacteria. *Journal of Molecular Biology* **428,** 2259–2264 (2016).

27. McPherson, A. *et al.* Introduction to protein crystallization. *Acta Crystallographica Section F Structural Biology Communications* **70,** 2–20 (2014).

28. Cooper, D. R. *et al.* Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta crystallographica. Section D, Biological crystallography* **63,** 636–45 (2007).

29. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology* **268,** 209–225 (1997).

30. Kuhlman, B. *et al.* Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. en. *Science* **302,** 1364–1368 (2003).

31. Gray, J. J. *et al.* Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *Journal of Molecular Biology* **331,** 281–299 (2003).

32. Kortemme, T. *et al.* Computational redesign of protein-protein interaction specificity. *Nature structural & molecular biology* **11,** 371–379 (2004).

33. Labonte, J. W., Adolf-Bryfogle, J., Schief, W. R. & Gray, J. J. Residue-centric modeling and design of saccharide and glycoconjugate structures. *Journal of Computational Chemistry* **38,** 276–287 (2017).

34. Das, R. Atomic-Accuracy Prediction of Protein Loop Structures through an RNA-Inspired Ansatz. *PLoS ONE* **8,** e74830 (2013).

35. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13,** 3031–3048 (2017).

36. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation* **12,** 6201–6212 (2016).

37. Leaver-Fay, A. *et al.* Scientific benchmarks for guiding macromolecular energy function improvement. *Methods in Enzymology. Methods in Protein Design* **523** (ed Keating, A. E.) 109–143 (2013).

Chapter 2

CAPRI

Adapted from Marze NA*, Jeliazkov JR*, Roy Burman SS, Boyken SE, DiMaio F, and Gray JJ, "Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28–35." *Proteins* **85(3)**, 479–486 (2017), with permission from the publisher. *Equal-contribution authors

2.1 Overview

The Critical Assessment of PRotein Interactions (CAPRI) is a continuous, community-wide evaluation of the performance of computational methods predicting the structure of protein complexes. I was a member of the Gray lab CAPRI team from 2014–2019, participating rounds 28–47. During my tenure, I observed a trend in CAPRI targets. From my perspective as a developer of modeling methods, targets could be categorized as either "easy", *i.e.*, targets that were typically useful for determining the accuracy of preexisting tools, or "challenging", which constitute targets that often necessitated the development of novel approaches. In this chapter I highlight several rounds featuring challenging targets that led to the development of new computational methods. Particularly, Targets 123, 124, and 160 featured camelid antibodies, which gave rise to the development of new homology modeling and docking methods; Targets 98–101 (among others) featured potentially disordered termini and protein regions, which led to an investigation of how best to model such regions; and Targets 104 and 105 (among others) featured highly solvated

interfaces, which inspired analysis, and later design, of crystallographic protein–protein interfaces.

2.2 Introduction

Proteins play important roles in cellular structure, metabolic activity, biochemical signaling, and multitudes of other biological functions. A protein's function is determined by its three-dimensional structure, particularly how this structure interacts with other proteins or other biological molecules to form complexes. Consequently, if the structure of protein complexes can be predicted, the nature of their function can likewise be elucidated. Though experimental methods exist to determine protein structure (X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy among others), these are costly, time consuming, and low-throughput. Computational structure prediction is an alternative that can quickly and cheaply generate a structural model of a protein complex.

The CAPRI competition offers an opportunity to evaluate the performance of state-of-the-art computational protein–protein docking methods[1]. A set of experimentally determined protein complex structures are withheld before publication, and protein docking groups are invited to submit their computational predictions of these structures. These predictions are assessed for accuracy by comparison with the experimentally determined structures. Thus, CAPRI serves as an important benchmark to evaluate the state of the field of computational protein docking, and to reveal remaining challenges. Our group has participated in CAPRI since its inception to evaluate the development of our docking method, RosettaDock[2]. RosettaDock is, at its core, a Monte-Carlo based rigid-backbone docking method with side chain optimization. RosettaDock is extensible, and several ancillary protocols have proven effective in previous CAPRI rounds[3]. The conformer-selection protocol EnsembleDock[4] and the flexible-loop induced fit protocol SnugDock[5] are among the most broadly useful.

2.3 Camelid targets have difficult-to-model H3 loops

Targets 123, 124 (both April 2017), and 160 (March 2019) each featured at least one camelid (single-chain) antibody[6]. Target 123 is the complex structure of a camelid-derived chaperone nanobody (called nb02) and the PorM N-terminal domain; PorM is an inner membrane protein involved in the Type IX secretion system (T9SS or PorSS)[7]. Akin to Target 123, Target 124 is the complex structure of another nanobody (called nb130) and the dimeric form of the PorM C-terminal domain. Finally, Target 160 is the assembly domain of a bacterial surface layer protein, with six distinct structural sub-domains, in complex with two chaperone nanobodies.

To model the nanobodies, I adapted the latest RosettaAntibody approach[8]. The nanobody sequences were deconstructed into four structurally conserved regions using regular expression to identify sequence motifs. These regions were the heavy-chain framework, capturing the conserved β-sandwich structure on which the loops rest, and the three complementarity determining regions (CDRs), comprising two loops with canonical structure (termed the CDR H1 and H2)[9] and one highly variable loop (termed the CDR H3)[10]. The sequence for each structural region was aligned against a database of sequences with known structure using BLAST+, and the homologous structures were grafted together following threading of the target sequence. Finally, to account for the highly variable nature of the CDR-H3 loop, I generated 1,000 models using *de novo* loop modeling methods[11].

We[i] utilized two approaches to modeling the nanobody partners. Structures were unavailable for the PorM N-terminal and C-terminal domains, so we used both homology and *de novo* modeling through the Robetta server[12], and *de novo* modeling with the *ab initio* protocol within Rosetta[13], combined with constraints derived from sequence co-evolution analysis[14]. The *ab initio* models of the PorM C-terminus had to be additionally docked symmetrically[15,16], as the structure was reported to be a dimer. For the surface layer

[i]Dr. Shourya Sonkar Roy Burman assisted in the modeling of the PorM termini.

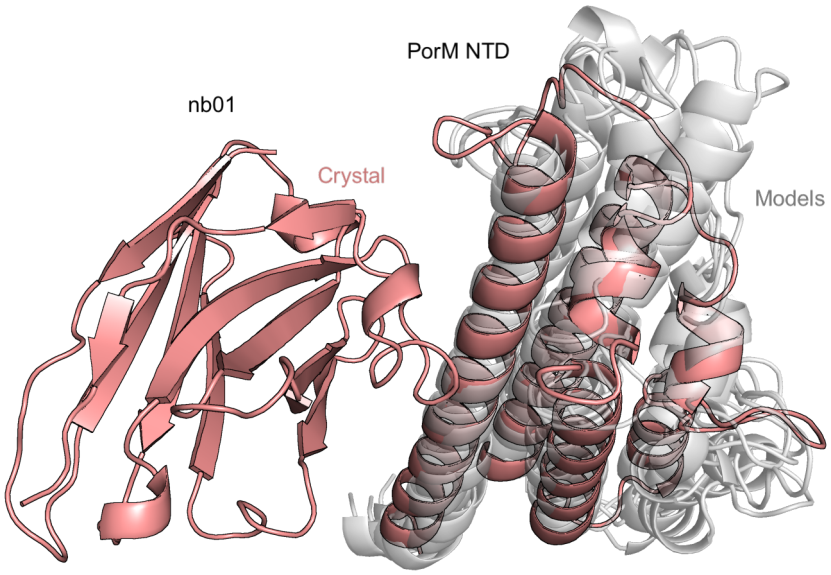

Figure 2.1: PorM N-terminal domain models (five homology models and five *de novo* models) aligned to the crystal structure (pink) of the PorM N-terminal domain in complex with nb01 reveal accurate modeling of four α-helical bundle domain (models in white). The model backbone RMSDs range from 2.7 Å to 4.5 Å, with most variability occurring in the helix-loop-helix motif (residues 159–189). Note that this comparison is not for the CAPRI Target 123 structure, but rather a related complex containing the same PorM N-terminal domain, but a different nanobody.

protein, the six sub-domain structures were given, but not the relative orientations. Thus, we[ii] separately considered the problems of identifying nanobody-domain interactions and assembling the domains.

For all targets, the best scoring individual models were globally docked with their partner using ClusPro[17–20]. Top-scoring models from ClusPro were further refined by SnugDock, a variant of RosettaDock which intercalates refinement of the CDR-H2 and -H3 loops with the standard approach.

For Target 160, the structures were not released as of this publication and thus could not be analyzed. For Target 123, only a structure of the PorM N-terminal domain in complex with a different nanobody was released as PDB ID 6EY0[7]. The structure (Figure 2.1) revealed that we had accurately modeled most of the N-terminal domain, particularly the region binding nb01 antibody. Most of our PorM N-terminal domain models missed some element

[ii]Dr. Sai Pooja Mahajan, Dr. Sudhanshu Shanker, and Ameya Harmalkar aided in the modeling of surface layer protein.

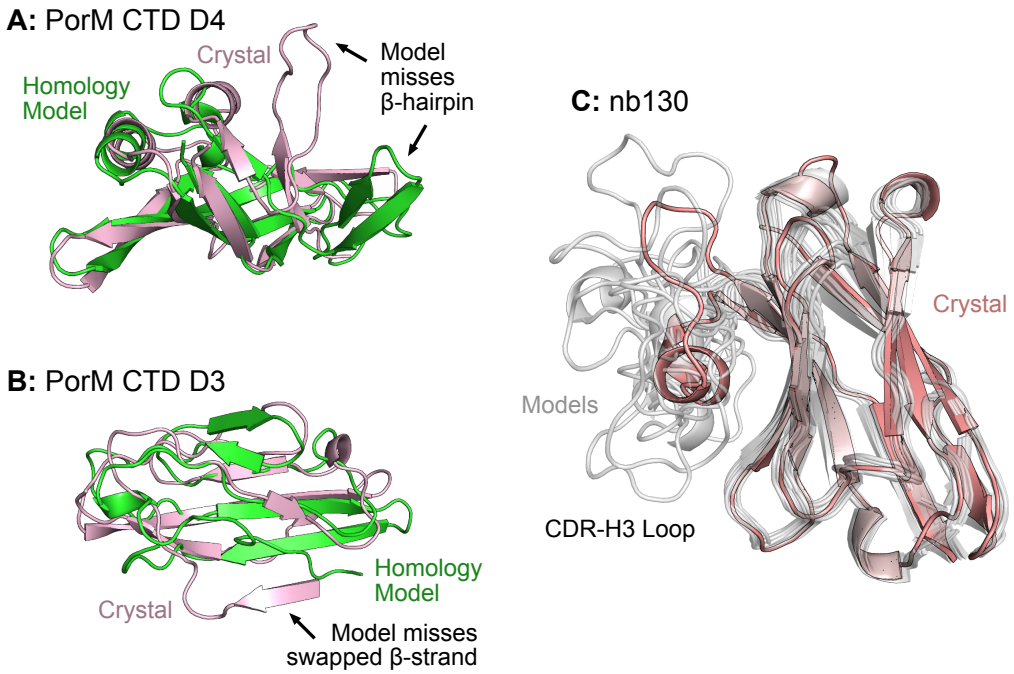

Figure 2.2: A comparison of PorM CTD sub-domains and their models, and nb130 and its models. The Porm CTD structure is decomposed into two sub-domains (D4 & D3, shown above in A & B respectively) as no accurate models were produced for the CTD dimer. Modeling the dimer was exceptionally challenging as it involved a β-strand swap between neighboring D3 domains. Additionally, modeling the linker between the D3 and D4 domains proved to be a challenge. The individual domains were modeled with reasonable accuracy (both at 4.9 Å backbone RMSD to native). The lowest-RMSD model for the D4 domain (A) missed the orientation of a mid-domain β-hairpin, whereas the lowest-RMSD model of the D3 domain (B) missed the β-strand that is swapped in the dimer. The ten lowest-scoring nanobody models range in backbone RMSD from 4.7–2.8 Å, with most of the difference occurring in the CDR-H3 loop (C). The native loop has two α-helical segments, which were challenging to model, that gave rise to a compact conformation. All but one model failed to capture both α-helical segments and be comparably compact.

of an internal helix-loop-helix motif (residues 159–189), with the lowest backbone RMSD model (at 2.7 Å) accurately capturing the helices, but not the linking loop.

Target 124, the nanobody-bound PorM C-terminal domain dimer complex, was released as PDB ID 6EY6[7]. The structure revealed that the PorM C-terminal domain consisted of two sub-domains (termed D3 and D4) and formed a homodimer with neighboring D3 sub-domains swapping β-strands across their β-sheets and a nanobody bound at the homodimeric interface. Despite modeling the D3 and D4 sub-domains to $\sim$5 Å backbone RMSD (Figure 2.2A, 2.2B), PorM C-terminal domain modeling failed. We were unable to correctly predict the D3–D4 orientation because we treated the two as a single domain

during our approach and did not accurately model the linker, resulting in significant lever-arm effects and poor input models for symmetric docking. Not to mention that we did not capture the strand swap. The nanobody was challenging to model due to its 21-residue CDR-H3 loop, which is well beyond the limits of accurate loop modeling[21]. Figure 2.2C shows that none of the modeled CDR-H3 loops were as compact as the crystallized nanobody, nor did they feature as many helical residues. In conjunction, the aggregate inaccuracies in modeling the PorM C-terminal domain dimer and the lengthy CDR-H3 loop contributed to our failure to model the complex correctly.

2.4 Flexible targets provide a sampling challenge

Targets 98–101 (December 2014) provided a combinatorial docking challenge which asked us to dock deubiquitinating enzyme UCH-L5, with or without its conjugate ubiquitin (Ub), to either of two inhibitors, RPN13 or INO80G. Unbound structures of UCH-L5, RPN13, and Ub were available (3IHR, 2KQZ, and 1UBQ, respectively). We homology modeled INO80G by threading from PDB structure 2KQZ, loop-building, and refining in Rosetta. Additionally, we built a homology UCH-L5–Ub complex by aligning the two proteins to PDB structure 4IG7. Using the FloppyTail protocol[22], we modeled the tails of RPN13, which are unresolved in 2KQZ, and the homologous regions of INO80G. We found no biochemical data or homology complexes that clearly identified a binding site, necessitating a global docking search. Due to the uncertainty in the monomer structures, we ran EnsembleDock with 30-member ensembles (generated by relaxing our top homology models). 20,000 decoys were generated for each target.

The assessment reveal that these targets were quite difficult: across all four targets, no CAPRI group submitted a model of acceptable-quality or better. Comparison of the complex binding mode to the unbound structures revealed that RPN13 undergoes a significant conformational shift upon binding, in which a helical bundle hinges open to bind around a helical element from UCH-L5, which itself undergoes a substantial kinking upon binding

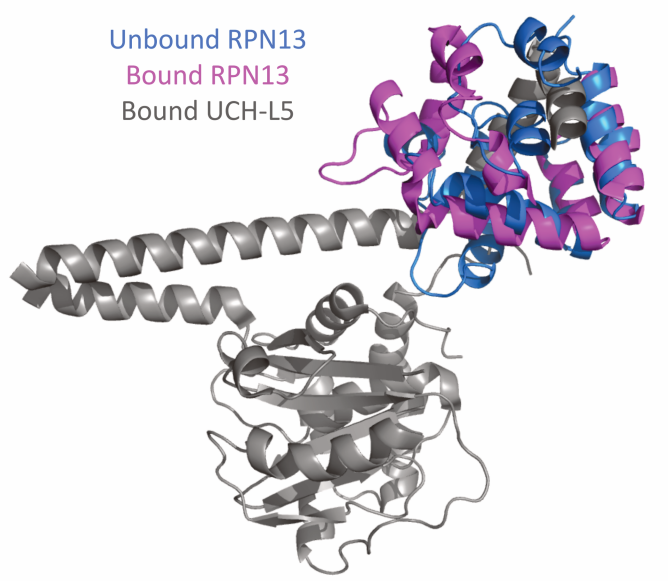

Figure 2.3: The bound UCH-L5–RPN13 complex (grey and pink), with the unbound RPN13 superimposed on top (blue). Upon binding, the RPN13 helical bundle hinges open to accommodate the UCH-L5 C-terminal helix.

(Fig. 2.3). Though INO80G has no unbound structure to compare with its bound forms, the inhibitor is similarly entwined with the UCH-L5 helix. This binding mode is doubly difficult to predict. Firstly, predicting conformational change upon binding has been observed to be difficult in previous CAPRI challenges[3], particularly when the change is so large. Secondly, the degree of structural entwinement between the two partners requires a hybrid folding/docking algorithm to predict correctly: the bound forms of RPN13 and INO80G would have high energies in solution due to their open hydrophobic pocket, and even if these forms could be predicted, due to the high degree of entwinement they would be almost impossible to dock by rigid-body methods.

2.5 Rosetta can position waters accurately at solvated interfaces

Targets 104 and 105 (March 2015) presented a dual challenge: first, to predict the complex structure of a DNase (PyoAP41 or PyoS2) with its cognate immunity protein (ImAP41 or ImS2), then to predict the mediating waters and side chains at the protein interface. While the DNase proteins had crystal structures available in the PDB, we had to generate

homology models for ImAP41 and ImS2. The available homologous structure was colicin Im2 (chain A in PDB ID 3U43), a previous CAPRI Target (47) with 50% identity to ImAP41 and 59% identity to ImS2. The starting complexes were then generated by aligning the homology models (ImAP41 or ImS2) and structures (PyoAP41 or PyoS2) to their homolog's position in PDB 3U43. For Target 104, we then used structural ensembles of PyoAP41 to account for a flexible loop at the interface and ran an local EnsembleDock to optimize the complex (50,000 decoys). For Target 105, we ran a local RosettaDock to optimize the complex (20,000 decoys). We used a new method for interface water predictions: HBNet with Bridging Waters (HBNetBW).

We expanded HBNet, a method for designing hydrogen bond networks[23], to include a statistical potential to capture water molecules that form bridging hydrogen bonds between side chains. The two-term potential utilizes the distance between the two protein atoms that hydrogen bond to the water molecule (acceptor or donor polar hydrogen) and the dihedral angle between those two atoms and their base atoms (e.g. the base atom for a carbonyl oxygen acceptor would be the carbon it is double bonded to, and the base atom of the polar hydrogen would the heavy-atom donor that it is covalently bonded to). We calibrated the potential using interface waters from the Top 8000 dataset[24] and bicubic spline interpolation; the two-dimensional function that defines the bridging water score is:

$$\text{score}(a_1, a_2, a_3, a_4) = f(\text{distance}(a_2, a_3), \text{dihedral}(a_1, a_2, a_3, a_4))$$

where a_2 and a_3 are the protein atoms hydrogen bonded to the bridging water, a_1 is the base atom of a_2, and a_4 is the base atom of a_2. To identify water positions during HBNet search, if two rotamers have a bridging water score below a specified threshold, they are connected as part of a potential hydrogen bond network and an explicit water molecule is placed at ideal geometry relative to the hydrogen-bonding atoms. We ran HBNetBW on each docked backbone, sampling rotamers of the interface residues to identify the most

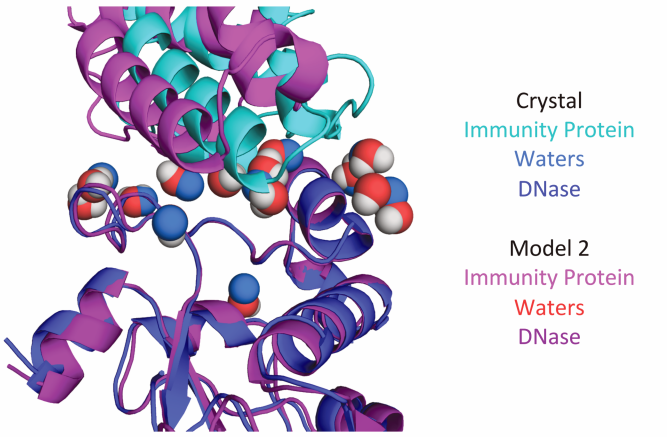

Figure 2.4: Our best medium-quality model for Target 105, superimposed with the crystal structure. Our model is colored in red/purple shades, while the crystal structure is colored in blue shades.

satisfied networks. There is a substantial energetic penalty associated with burying polar atoms that do not participate in hydrogen bonds (either to solvent or other protein atoms); thus, we hypothesized that using this criterion would be advantageous for discriminating between docked complexes.

For Target 104, all of our models were incorrect. An *ex post facto* analysis revealed that our homology model had the correct complex orientation, and that our docking simulation moved the complex away from that conformation. For Target 105, all four of our submitted models were of medium quality, the best having an interface RMSD of 1.757 Å and recovering 48.1% of native interface contacts. Similar to Target 104, however, the unrefined homology model had a more native-like orientation than our docked model. One of our models from Target 105 had a fair-quality water prediction (Figure 2.4), recovering 11.8% of native interactions with waters, indicating that HBNet can be useful even without a perfectly-aligned interface.

After the CAPRI blind challenge, we ran HBNetBW on the revealed crystal structures for Targets 104 and 105 and the closest homology model to each. We removed water molecules from the structures. We then relaxed (cycles of minimization and side-chain repacking) the structures using Rosetta. Next, we ran HBNetBW using identical parameters to those during analysis of submitted docked complexes. In regions of the interface where the

backbone was close to that of the crystal structure, the native side-chain hydrogen-bond networks were largely recapitulated, and a couple of the bridging water molecules were placed in agreement with interface waters in the crystal structure; for example, running HBNetBW on the Target 105 homology model generated a network with a bridging water molecule between Tyr640, Tyr55, and His34 that is in close agreement to the experimental crystal structure. However, many false-positive networks and water placements were also generated – multiple networks are identified for each fixed-backbone decoy, making it challenging to choose which networks and water placements to keep and which to discard. Ranking networks according to satisfaction and connectivity led to success in designed protein-only networks[23]; however, as used here, these metrics are only as reliable as the bridging water identification and placement, and our results suggest that there is significant room for improvement to both.

2.6 Discussion

Over the years, challenges presented by CAPRI targets have led to the development of fully-fledged modeling protocols. For example, following the task of modeling a lysozyme–inhibitor complex (which is relevant in a low pH) target, dynamic residue protonation was introduced to RosettaDock[25]. In another example, antibody–complex targets led to the development of SnugDock, which introduced refinement of loops during docking[3]. The targets highlighted in the chapter identified future research directions in modeling camelid antibodies, flexible regions, and solvated interfaces, guiding the direction of my PhD research.

In the process of modeling the three camelid antibody targets, I identified shortcomings in our methodology. First, the grafting step of the RosettaAntibody protocol lacked the capacity to assemble a heavy-chain only antibody, due to the assumed presence of a light chain. Second, the database for grafting lacked templates sourced from camelid antibody structures and, owing to its manual curation, had not been updated in years. Third, as with

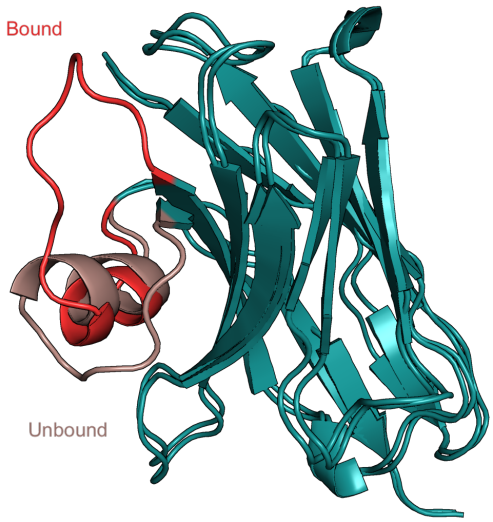

Figure 2.5: A comparison of the bound and unbound crystal structures of nanobody 06 exemplifies the large motions associated with flexible loop regions. The CDR-H3 loop (salmon [PDB ID 5E7B] or dark red [PDB ID 5E7F]) partially unfolds a helix and shifts from interacting with the framework β-sheets to the interacting with the other CDRs.

grafting, Rosetta SnugDock could not natively dock a single-chain antibody to an antigen. Finally, the CDR-H3 loop accuracy of models tended to be low due to its inherent flexibility. Alleviating these deficiencies (among others) is the basis of Chapter 3.

The difficulty of accurately modeling flexible regions was observed with Targets 98–101, in addition to the CDR-H3 loops. Modeling these regions was particularly challenging due to the size of the possible conformation space and, often, the large structural re-arrangement between between the initial and final conformation (as showcased in Figures 2.3 and 2.5). From these observations arises the need to more thoroughly characterize the motions of flexible regions and the biological consequences of flexibility. In Chapter 5, I systematically investigated the large conformational freedom and energy landscape accessible to the disordered termini of the Hfq protein across several bacterial species and its effects on Hfq's RNA annealing function. In Chapter 4, I assessed the differences in CDR-H3 loop flexibility between naïve and antigen-experienced antibodies by combining Rosetta loop modeling methods and a graph-theoretical approach for estimating atomic rigidity.

Finally, we observed unexpected success in modeling water positions at a highly sol-

vated interface, even when the interface itself was not accurately modeled. This was surprising because Rosetta was developed to model *ab initio* folding and optimized in many aspects for this task, including the sampling strategies and scoring potentials – that is to say hydrophobic interactions tend to be more emphasized and better modeled than electrostatic interactions – and Rosetta, by default, uses an implicit solvent model. In this context, the accurate placement of water molecules at an interface was a significant achievement. This modeling success inspired me, in part[iii], to pursue the ambitious design of highly solvated protein–protein interactions at crystallographic interfaces, detailed in Chapter 6.

[iii] Additional motivation was provided by the broad potential impact of a reliable crystal design tool.

References

1. Vajda, S., Vakser, I. A., Sternberg, M. J. E. & Janin, J. Modeling of protein interactions in genomes. *Proteins* **47,** 444–6 (2002).
2. Gray, J. J. *et al.* Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *Journal of Molecular Biology* **331,** 281–299 (2003).
3. Sircar, A., Chaudhury, S., Kilambi, K. P., Berrondo, M. & Gray, J. J. A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19. *Proteins: Structure, Function, and Bioinformatics* **78,** 3115–3123 (2010).
4. Chaudhury, S. & Gray, J. J. Conformer Selection and Induced Fit in Flexible Backbone Protein–Protein Docking Using Computational and NMR Ensembles. *Journal of Molecular Biology* **381,** 1068–1087 (2008).
5. Sircar, A. & Gray, J. J. SnugDock: Paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Computational Biology* **6,** e1000644 (2010).
6. Muyldermans, S. Nanobodies: Natural Single-Domain Antibodies. *Annual Review of Biochemistry* **82,** 775–797 (2013).
7. Leone, P. *et al.* Type IX secretion system PorM and gliding machinery GldM form arches spanning the periplasmic space. *Nature Communications* **9,** 429 (2018).
8. Weitzner, B. D. *et al.* Modeling and docking of antibody structures with Rosetta. *Nature Protocols* **12,** 401–416 (2017).
9. North, B., Lehmann, A. & Dunbrack, R. L. A new clustering of antibody CDR loop conformations. *Journal of Molecular Biology* **406,** 228–256 (2011).
10. Weitzner, B. D. & Gray, J. J. Accurate Structure Prediction of CDR H3 Loops Enabled by a Novel Structure-Based C-Terminal Constraint. *The Journal of Immunology* **198,** 505–515 (2016).
11. Stein, A. & Kortemme, T. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS ONE* **8** (ed Zhang, Y.) e63090 (2013).
12. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods in enzymology* **383,** 66–93 (2004).
13. Raman, S. *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics* **77,** 89–99 (2009).

14. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residueâĂŞresidue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences* **110,** 15674–15679 (2013).

15. Andre, I., Bradley, P., Wang, C. & Baker, D. Prediction of the structure of symmetrical protein assemblies. *Proceedings of the National Academy of Sciences* **104,** 17656–17661 (2007).

16. Burman, S. S. R., Yovanno, R. A. & Gray, J. J. Flexible backbone assembly and refinement of symmetrical homomeric complexes. *bioRxiv* (2018).

17. Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics (Oxford, England)* **20,** 45–50 (2004).

18. Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic acids research* **32,** 96–9 (2004).

19. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function and Genetics* **65,** 392–406 (2006).

20. Kozakov, D. *et al.* How good is automated protein docking? *Proteins* **81,** 2159–66 (2013).

21. Ó Conchúir, S. *et al.* A Web resource for standardized benchmark datasets, metrics, and rosetta protocols for macromolecular modeling and design. *PLoS ONE* **10** (ed Zhang, Y.) e0130433 (2015).

22. Kleiger, G., Saha, A., Lewis, S., Kuhlman, B. & Deshaies, R. J. Rapid E2-E3 Assembly and Disassembly Enable Processive Ubiquitylation of Cullin-RING Ubiquitin Ligase Substrates. eng. *Cell* **139,** 957–968 (2009).

23. Boyken, S. E. *et al.* De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science (New York, N.Y.)* **352,** 680–7 (2016).

24. Chen, V. B. *et al.* MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66,** 12–21 (2010).

25. Kilambi, K. P., Reddy, K. & Gray, J. J. Protein-Protein Docking with Dynamic Residue Protonation States. *PLoS Computational Biology* **10,** e1004018 (2014).

Chapter 3

RosettaAntibody and SnugDock Development

3.1 Overview

As demonstrated in Chapter 2 and other prior work[1], the most scientifically challenging aspects of antibody modeling and antibody–antigen docking with RosettaAntibody and Rosetta SnugDock are (1) modeling the CDR H3 loop, (2) modeling camelid antibodies, and (3) identifying templates for exotic antibodies[i]. Additionally, non-scientific issues arise for both end users and developers of the software. End users are burdened by an unwieldy options system and unclear default settings, whereas developers at times must interact with poorly structured code. In this chapter, I report advances I have made towards addressing these challenges. I contributed several direct changes to address the scientific questions: (1) I implemented an updated loop modeling implementation in Rosetta Antibody and SnugDock that simplified the code structure and permitted for the use of new loop modeling techniques such as loop hash or fragment insertion; (2) I enabled the modeling of camelid antibodies throughout both methods by relaxing assumptions about the presence/absences of chains in the antibody and implementing a novel FoldTree in SnugDock; (3) I developed a program to automatically update the template database for RosettaAntibody. To rigorously test the effects of my changes, I contributed scientific

[i]For example, antibodies that have atypical CDR lengths or sequences, or come from species other than mouse or human.

benchmarks. Then, I simplified the command-line options and set reasonable defaults to make our software more easily accessible to end users. Finally, with others[ii], I re-factored the grafting step of RosettaAntibody to be more sustainable and developer friendly.

3.2 Introduction

RosettaAntibody is a hybrid modeling tool, utilizing both homology and *de novo* modeling to predict the structures of antibodies from sequence. The approach is inspired by observations of variance in antibody crystal structures and can be divided into two stages. The first stage, referred to as "grafting", selects templates based on homology for conserved structural regions of the antibody (the CDR loops, the framework regions [FR], and the relative orientation) and assembles a single model from the multiple sourced templates (Figure 3.1). The second stage, referred to as "refinement", *de novo* models the CDR-H3 loop by random perturbation and kinematic closure[2] and refines the heavy-chain–light-chain (V_H–V_L) relative orientation by a generic local docking approach[3]. In a typical simulation, the user, with a FASTA file as the only input, will first generate ten grafted models, each differing only in the V_H–V_L relative orientation, then further generate 1,000 refined models for the most likely orientation and 200 models for the remaining orientations, resulting in 2,800 models of which (usually) the ten lowest scoring are selected as those most likely to approximate the native structure.

While RosettaAntibody has proven itself to be a useful tool, antibodies are often not modeled in isolation. Rather, the key biological interest lies in how a given antibody might interact with its cognate antigen. This question is addressed by Rosetta SnugDock. Given a plausible initial conformation of an antibody–antigen complex, SnugDock simulates the antibody–antigen interaction by simultaneously optimizing the antibody–antigen interface, the V_H–V_L interface, and the six CDR loops, with refinement of the CDR-H3 and -H2 loops. In a typical simulation, the user will provide a starting conformation along with

[ii]Sergey Lyskov and Dr. Brian D. Weitzner.

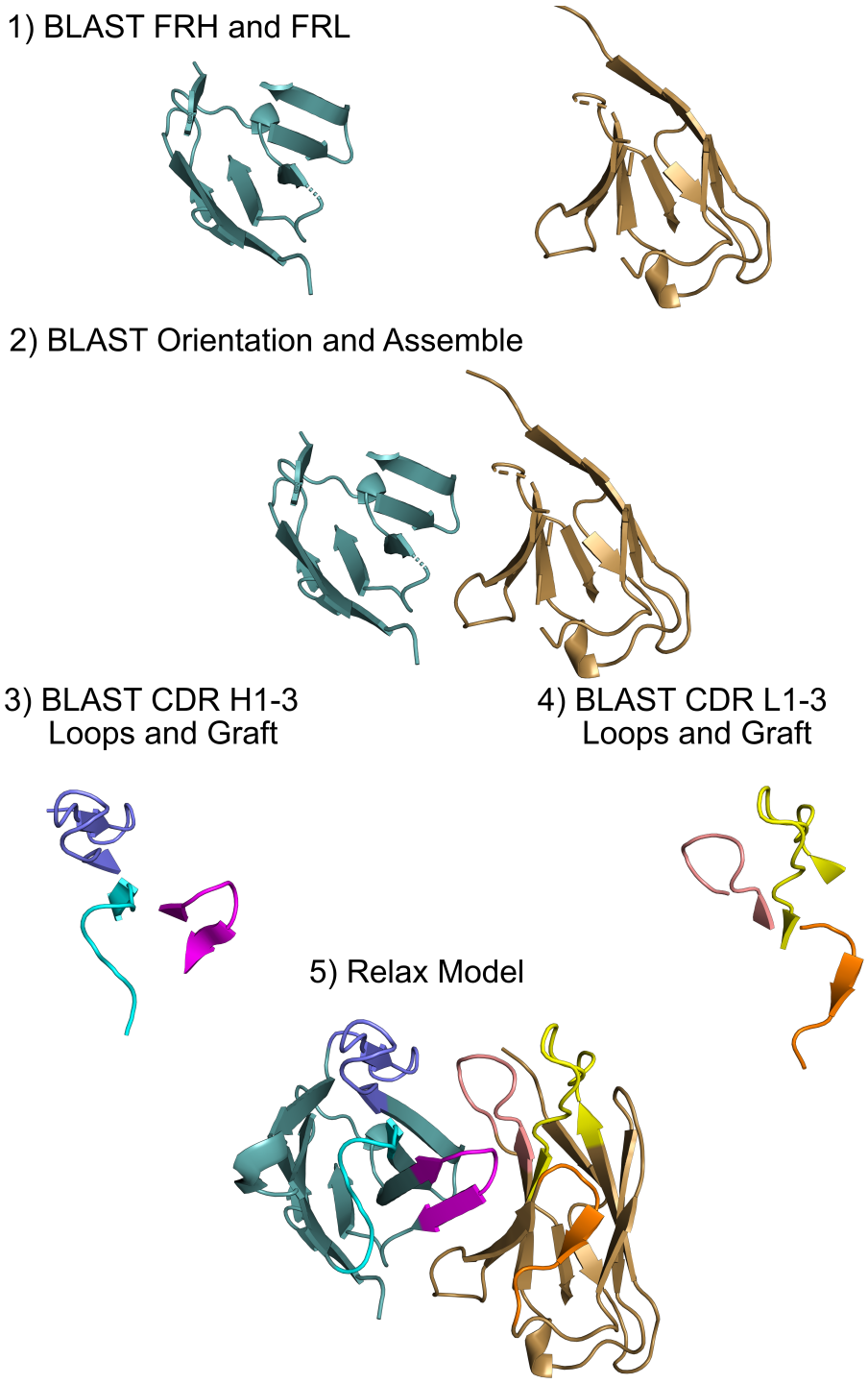

Figure 3.1: Schematic of the RosettaAntibody grafting step.

an ensemble of antibody models (or structures) and an ensemble of antigen models (or structures) to SnugDock, generate 1,000 docked models, and select the 10 lowest-scoring models as the ones most representative of the native structure.

In recent years, most usage of these modeling tools has been coupled. A frequent modeling task for users has been to first model a set of antibody sequences (known to bind a target antigen), and then dock the antibody models to the antigen (for which the structure may be known or modeled). In one such example, these antibody sequences were acquired from sequencing following phage display against a target implicated in Celiac disease pathology[4]. There are many possible sources of antibody sequences, including high-throughput sequencing studies that can produce on the order of 10^5 paired V_H–V_L sequences[5]. It is not unreasonable that antibody sequence-determination methods will continue to improve and generate more data in the near future, thus it is expected that the need for accurate computational modeling will similarly rise[6]. In anticipation of these trends, I have contributed multiple improvements to the RosettaAntibody and SnugDock protocols that enable sustainable future development.

3.3 Making antibody grafting object-oriented

Before expanding the functionality of Rosetta Antibody, Dr. Brian D. Weitzner, Sergey Lyskov, and I improved the stability of the grafting stage. We refactored the old approach, contained entirely by the Python script `graft.py`, into object-oriented, C++ code that comprised multiple objects and classes, united in a single grafting application called `antibody.cc`. The re-factored application split the tasking of antibody grafting into three sub-tasks: (1) identifying structural regions for grafting, (2) identifying templates for the regions, and (3) assembling the templates into a single model. Unified modeling language (UML) diagrams of the classes central to the re-factored grafting approach are included as Supplemental Figures 3.A.1–3.A.3.

In `antibody.cc`, an input sequence is used to construct an `AntibodySequence` object,

Table 3.1: Antibody structural regions used in RosettaAntibody and corresponding numbering under the Chothia convention. These definitions are hybrid Chothia/Kabat. FRH/L do not complement CDR sequences as there are additional (non-CDR) loops that should be excluded when selecting a template. *What exactly contributes to V_H–V_L orientation is unclear, so the combined sequence is used for comparison.

Region	Definition
CDR L1	24–34
CDR L2	50–56
CDR L3	89–97
FRL	10–23 35–39 46–49 57–66 71–88 98–104
CDR H1	26–35
CDR H2	50–65
CDR H3	95–102
FRH	10–25 36–39 46–49 66–94 103–109
Orientation*	L5–L104 H5–H109

which is a `struct` containing a `std::string` for each structural component of the antibody. We define eight structural components: the six CDRs (three from each chain) and the two frameworks (one from each chain). Our definitions are compared to the canonical Chothia numbering scheme in Table 3.1. Detection of these regions is done either by regular expressions (with the `Regex_based_CDR_Detector` class) or by external specification through a JSON-formatted file (with the `Json_based_CDR_Detector` class). The `AntibodySequence` object is written such that any class can pass structural definitions to it.

Next, a "structural component selector" (SCS) is instantiated and configured. Currently, we use `SCS_BlastPlus` which identifies templates by a BLAST+[7] comparison against a pre-constructed database. The code is structured such that one could use any method for selection, as long as it inherits from the correct base class: `SCS_LoopOverSCs`. For example, Dr. Brian D. Weitzner has implemented a selector using custom substitution matrices rather

than the default ones (*e.g.* PAM32) supplied by BLAST[iii]. Following the identification of possible templates, the most favorable ones must somehow be selected. This is done by a filtering and sorting step. For BLAST results, the following filters are currently implemented (as `SCS_BlastFilter_by_XYZ`):

- sequence length,
- alignment length,
- sequence identity,
- template resolution,
- outlier (read from external list),
- template B-factor,
- OCD[iv], and
- template PDB ID.

Finally, the SCS results are sorted by bit score and resolution. This could be altered in the future by replacing or altering the `SCS_BlastComparator_BitScore_Resolution` class.

After sorting, the final stage of grafting is assembling the individual structural components into a single model. This is handled by the `graft_cdr_loops` function, which assembles antibodies in the following order. First, the templates for the heavy and light frameworks are loaded. Next, the framework query sequences are threaded on to the corresponding templates and the threaded frameworks are aligned to the template orientation. Finally, this is followed by sequential grafting and threading of each CDR on the assembled frameworks. Grafting is done by the `AntibodyCDRGrafter` class developed by Jared-Adolf Bryfogle for antibody design and detailed elsewhere[9]. This final stage can be repeated multiple times, with multiple ranked sequence alignments producing a set of models. In a standard antibody homology modeling problem, one should produce models with ten

[iii] However, this and any future selector would have to be called in `antibody.cc`, which is not currently the case.

[iv] Orientation Coordinate Distance is a measure of V_H–V_L relative orientation, originally defined by my colleague Dr. Nick A. Marze[8].

different orientations to hedge against the challenge of correctly predicting the relative V_H–V_L orientation[8].

3.4 Automating the template database

A homology modeling method is only as good as the structural database it relies on. A method relying on a database with few structures is unlikely to produce good models. The CAPRI assessment (Chapter 2) revealed the outdated and rigid nature of the RosettaAntibody template database: when tasked with modeling camelid antibodies, we had struggled to find good templates. Despite the knowledge of hundreds of camelid antibody structures at the time, none had permeated the database. Why? Prior to improvements reported below, the template database was constructed manually.

As I never participated in the manual construction of the original database[10], I cannot neither speculate on the rationale and reasoning behind it nor its evolution throughout the years. At the onset of my tenure, the following files were essential to the antibody database:

- Chothia-numbered PDB structures,
- BLAST database[v], categorized by region and length (if the region is a loop),
- a file containing B-factors[vi],
- a file containing the OCD between all antibodies in the database,
- a file containing outliers, and
- a summary file with CDR and FR sequences for all antibodies, "antibody.info".

Based on the presence of these files, I wrote a script to automatically update each. The PDBs are now downloaded directly from SAbDab[11] in Chothia-numbered format. The only requirement is that the PDB crystal structures have higher resolution than 3.0 Å. The PDBs are then subject to quality checks. Using PyRosetta[12], I check for missing residues or

[v] Intermediate "info" files were used to store the CDR sequences used for BLAST database construction, *e.g.* "cdr.info".

[vi] Actually, this file contained boolean values representing whether or not the structure met some criterion, but it is not clear exactly what the criterion was.

Table 3.2: Comparison between the last iteration of the manual database and the first iteration of the automatic database (February 15th, 2019). Template counts for each region are shown.

	Old Count	New Count	Overlap
All CDRs	1,902	2,611	1,560
FRH	1,785	2,390	1,427
FRL	1,577	2,832	1,111
Orientation	1,003	1,721	749

unrealistic peptide bond lengths in the regions crucial to modeling (*i.e.* CDRs and conserved framework residues), excluding any region if it does not pass the check. During the quality checks, sequences are extracted for each region and written to temporary "info" files. Then these files are read in, grouped by region and length, and output in FASTA format. The BLAST databases are then constructed as:

```
makeblastdb -in fasta -dbtype prot -title database.cdr.length -out
   database.cdr.length
```

Finally, the relative orientations coordinates for each antibody, the relative orientations for each pair, and the B-factors are extracted and written to files (to be used for filtering during the grafting step).

In addition to automating database updating, I have also written comparison scripts to assess the changes in the database from update to update and developed a scientific benchmark to evaluate any changes in grafting accuracy. Table 3.2 summarizes the changes following the shift from the manual to the automatic database. Shifting resulted in an increase in templates for all regions. PDBs shared by both databases were used as a comparison to test whether or not the auto-update script is producing reasonable output.

Table 3.3 highlights the differences at the sequence level following the shift. In general, these differences are minimal (on the order of $\sim$1%). The differences arise for a variety of reasons. In cases when multiple antibodies are present in the same PDB asymmetric unit, different antibodies are selected from the multiple possibilities. In other cases, PDBs are omitted in the new database due to geometry issues (missing atoms or non-ideal C–N distances) in critical regions. Finally, the most prevelant case for FRH or CDR-H2 loop

Table 3.3: Comparison of sequences extracted for identical regions for identical PDBs. There are (in terms of percent) very few mismatches.

	Mismatch Count	Mismatch Percent
CDR H1	30	1.1%
CDR H2	60	2.2%
CDR H3	33	1.3%
FRH	27	1.1%
CDR L1	11	0.4%
CDR L2	15	0.6%
CDR L3	8	0.3%
FRL	2	<0.1%
Orientation	10	1.3%

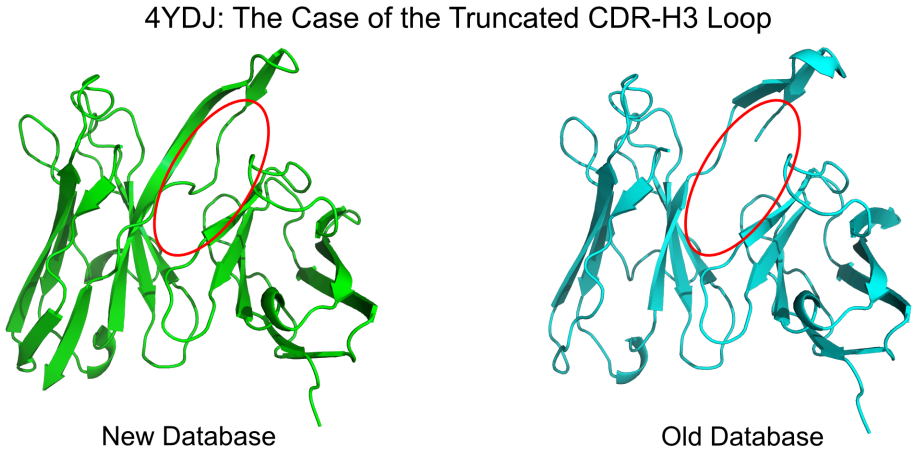

Figure 3.2: By importing Chothia-numbered PDBs, testing for the presence of all backbone atoms, and testing for good C–N bond geometry, the new database properly processes and retains the entire CDR-H3 loop for 4YDJ. The same PDB in the old database was (inexplicably) missing a large portion of CDR-H3 loop atoms.

mismatches is when the numbering schemes are not identical between database. It is unclear why the old database has some incorrectly numbered loops. It is possible that the old database may have used regular expression, for which failures would have been difficult to detect. In the new database, sequences are derived from Chothia-numbered PDB files, from a validated server with an error rate close to 0%[11,13]. Figure 3.2 shows the consequences of errant numbering: for this antibody (4YDJ), the CDR-H3 loop is missing atoms in the old database.

3.5 Modeling camelid antibodies with Rosetta

Exposure to the challenge of modeling camelid antibodies in Chapter 2, not only triggered a re-think about the database, but also revealed an inability to model camelid antibodies with RosettaAntibody and SnugDock. Interestingly, camelid antibody structure prediction was possible at one point[14], but that ability was lost due to underlying changes in the Rosetta code base[15]. While re-enabling camelid antibody modeling with RosettaAntibody was trivial, requiring only minor modifications to the grafting step, camelid antibody docking necessitated more substantial changes.

Enabling the docking of single-chain antibodies in SnugDock was challenging. In the course of a SnugDock simulation, both the relative orientation between the heavy and light chain and the orientation between the antibody and antigen are refined. This case constitutes a multi-body docking problem. Rosetta does not natively support multi-body docking. Thus, the SnugDock protocol had to make certain assumptions and alterations to the kinematic information stored during the simulation. The object storing this data is known as the FoldTree[16].

Briefly, the FoldTree is an object within Rosetta that dictates in what order residue/protein positions should be updated. In its implementation, the FoldTree is an acyclic graft comprising directed edges connected by directed jumps. Edges represent physically connected objects such as a single polypeptide chains. This representation is efficient and the position for a given residue can be determined if the dihedral angles of the preceding residues in the FoldTree are known[17]. Jumps represent virtually connected objects, such as two proteins. Again, this representation is efficient as the position of the second protein can be specified by a rotation matrix and a translation vector relative to the first protein. However, in this implementation only a single jump can be updated at time, making it impossible to dock multiple proteins relative to one another. If three protein chains (A, B, and C) were to be docked, one could only move a single chain relative to the others, in a

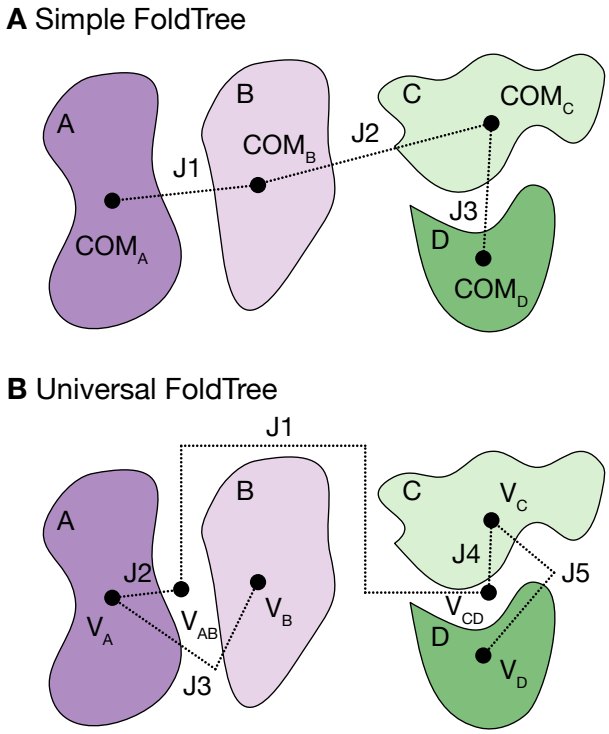

Figure 3.3: (A) An implemented example of a simple FoldTree for multiple body docking, here with four proteins. Proteins are shown as blobs and labeled A, B, C, and D. In this setup, A can dock to BCD, AB can dock to CD, and ABC can dock to D, but B could not move independently. (B) An implemented example of a Hierarchical FoldTree, as currently implemented in Rosetta SnugDock, here virtual atoms, placed at corresponding centers of mass, are shown as points and labeled. Relevant jumps are shown as dotted lines and labeled. Jumps connecting center of mass virtual atoms (V_A, V_B, V_C, and V_D) and the N-termini of corresponding polypeptide chains are omitted. In this configuration any protein within a complex can be docked to its neighbors and complexes can be docked to each other. For example, by docking across J1, the AB complex can be docked to the CD complex, or by docking across J5, protein C can be docked to protein D. Another equally valid approach would be to connect individual proteins to the complex center of mass (*i.e.* J5 would connect V_D to V_{CD} and not V_C).

given FoldTree (Figure 3.3A). For example, if one wanted to dock AB to C, and then dock AC to B, one would have to construct two separate FoldTree objects (one in which A jumps to B jumps to C, so C can move relative to the AB complex, and another where A jumps to C jumps to B, so B can move relative to the AC complex). The three proteins could not move simultaneously because storing the translations and rotations for all three proteins in the FoldTree would break the acyclic property.

SnugDock avoids this issue by separating the antibody–antigen docking simulation into a collection of smaller simulations, each with its own FoldTree[18]. However, assumptions

are necessarily made when handing off the collection of atoms from one simulation to the next (*e.g.* the first polypeptide is the heavy chain, the second is the light, and the third is the antigen), as the FoldTree object does not have any knowledge of the collection of atoms to which it is attached[vii]. Prior to my changes, it was impossible to input a single chain antibody to a SnugDock simulation because of the assumed presence of a light chain.

To correct this issue, I introduced what is known as a "Hierarchical FoldTree", originally proposed by Dr. Nick A. Marze in his PhD dissertation as a "Universal FoldTree"[19], to Rosetta SnugDock. This FoldTree places what are known as "virtual" residues at protein and complex centers of mass and then connects the polypeptide chains in a hierarchical fashion, such that complexes of interest are grouped together[viii] (*e.g.* the two antibody chains or any number of antigen chains). By using virtual atoms at the centers of mass, transformations and rotations between most potential docking partners are known. Thus, complexes can dock to other complexes and neighboring polypeptide chains within complexes chains can dock to each other (Figure 3.3B). An additional benefit of using virtual residues, rather real ones, is that each polypeptide chain can have its own internal FoldTree, which is then connected to the Hierarchical FoldTree. This permits FoldTree-dependent modifications within in each chain (such as loop modeling) to take place, without necessitating a new FoldTree. As a result, the updated SnugDock protocol uses a single FoldTree throughout the simulation.

3.6 Introducing new loop modeling approaches

Having the ability to model and dock camelid antibodies does not necessarily result in the ability to produce high-accuracy models. As was shown in Chapter 2, camelid antibody models suffer from significant inaccuracies in the CDR-H3 loop. In fact the CDR-H3 loop is the most challenging region to model for most antibodies[1]. CD3-H3 loop prediction is

[vii] Collections of atoms (typically proteins) are stored in Pose objects in Rosetta. It would be silly and inefficient to store the same information twice by replicating Pose data in the FoldTree.

[viii] In general, the design choice to connect neighboring chains is purely stylistic and an equally valid alternative is to connect the chains to the complex center of mass, so one can dock each chain against the complex.

a substantial challenge in the field[20]. The challenge arises from the immense diversity of the CDR-H3 loop, which occurs both at a sequence[21] and structural level[22]. One possible approach to improving loop modeling accuracy is to introduce and apply novel loop modeling methods. Since CDR-H3 loops are more diverse than standard protein loops, why should we use standard loop modeling approaches?

To that end, I have collaborated with Prof. Tanja Kortemme's lab at the University of California, San Francisco to implement new loop modeling methods in RosettaAntibody and Rosetta SnugDock. Her group's recent advances in loop modeling build on their development of kinematic loop closure (KIC)[2,23] by introducing the use of fragments to either perturb loop structure before closure (termed "fragment KIC") or to assist closure of a perturbed loop (termed "loop hash KIC"). Fragment KIC has been benchmarked on a diverse set of protein loops and shown to improve loop modeling performance (Xingjie Pan, unpublished). On the other hand, Loop hash KIC is still under development and testing on regular protein loops. Thus while both methods are implemented for antibody modeling and docking, I only assessed the performance of fragment KIC on the more challenging problem of antibody loops.

Figure 3.4 compares the distribution of minimum CDR-H3 loop RMSDs produced for a set of 49 antibodies. How the antibodies were selected and the simulations conducted is detailed in the following section, with sample command lines given in the Appendix. Fragments were selected using the Robetta server, which implements the fragment picker[24]. Briefly, the fragment picker protocol takes as input a protein sequence and breaks it into all overlapping windows of a particular length (three and nine in this case). For each window, secondary structure propensities are calculated from the sequence, and 200 fragments are selected from a pre-constructed database[ix] by comparing sequences and secondary structure propensities. To give an example, a 50 residue protein has $50 - 3 = 47$ windows and, for each, 200 fragments will be selected.

[ix] Termed Vall, the database contains non-redundant sequences with known structure in the PDB.

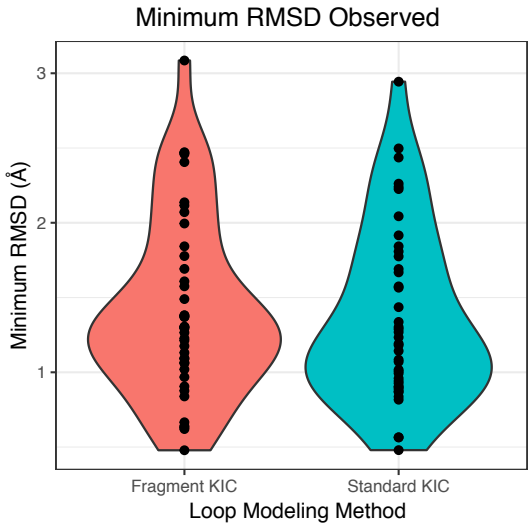

Figure 3.4: The distributions of the minimum CDR-H3 loop RMSDs observed for all antibodies in the benchmark, for two loop modeling methods, do not significantly differ according to Student's t-test (p-value = 0.67).

From Figure 3.4, it is clear that including fragments does not improve overall CDR-H3 loop prediction accuracy, as the minimum RMSD distributions are identical. This result is not surprising considering it has recently been reported that 30% of CDR-H3 loops do not have matching ($<$ 1Å RMSD) four-residue fragments in the PDB[22]. However, as the fragments used in my study were of length three or nine, I decided to quantify the structurally similarity of fragments and loops in both my antibody set and Xingjie's protein set.

To evaluate structural similarity, I compared every fragment to its overlapping loop segment. For overlapping residues, I calculated the difference between the fragment and loop backbone dihedral angles. I then expressed this difference as a chord distance: $D^2(\theta_1, \theta_2) = 2 - 2\cos(\theta_2 - \theta_1)$ and summed over the overlapping residues: $\langle D \rangle = \frac{1}{n}\sum_n (D^2_\phi + D^2_\psi)/2$. Thus, $\langle D \rangle$ has a minimum of 0, if a fragment matches a loop exactly, and a maximum of 4, if a fragment differs by 180 degrees at every dihedral from a loop.

The structural comparison of Rosetta-derived fragments and protein or antibody loops is shown in Figure 3.5 for three-residue fragments and Figure 3.6 for nine-residue fragments. These figures depict cumulative distributions of all fragment to all loop comparisons: for

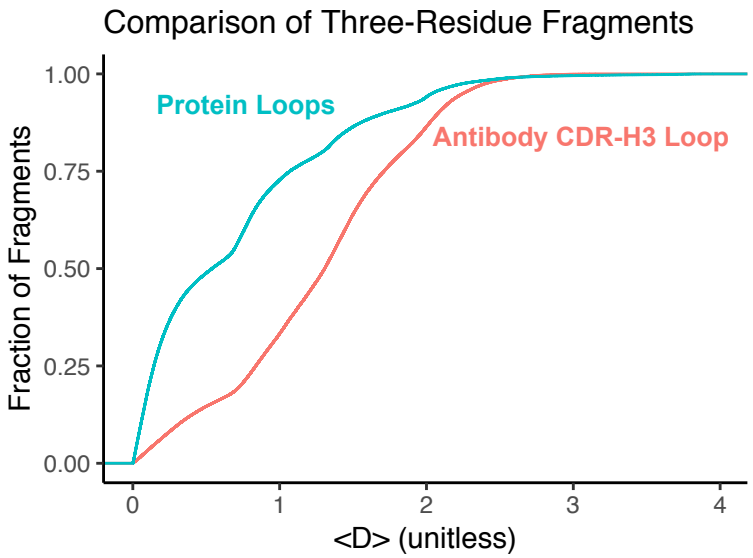

Figure 3.5: Three-residue fragments from the PDB are more structurally similar to protein loops than to the antibody CDR-H3 loop. The cumulative distribution function yields the probability (y-axis) that a fragment is within a certain chord distance (x-axis) of a loop.

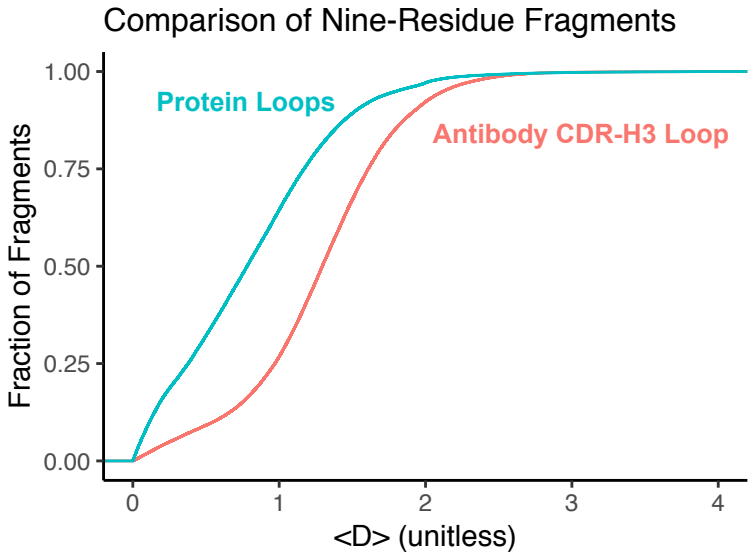

Figure 3.6: Nine-residue fragments from the PDB are more structurally similar to protein loops than to the antibody CDR-H3 loop. The cumulative distribution function yields the probability (y-axis) that a fragment is within a certain chord distance (x-axis) of a loop.

each $\langle D \rangle$ value along the x-axis, the y-axis represents the fraction of fragments with similar or smaller $\langle D \rangle$. Even at a quick glance, these figures confirm the earlier findings that CDR-H3 loops are structurally distinct, as over nearly all values of $\langle D \rangle$ protein loops have more matches in their respective fragment set.

3.7 Scientific tests

When developing software, it is important to track how changes to the underlying code affect software functionality. In the worst case, a developer would want to catch changes that break the software before releasing an unusable product to the public. In less severe cases, a developer would want to know if their changes improve or worsen the software in terms of some performance metric (*e.g.* speed or for scientific software accuracy). In Rosetta, changes in code functionality are tracked through a series of tests, each operating on a different scale.

At the smallest scale of object-oriented code, unit tests assess the individual classes and functions, by evaluating how the code processes pre-determined queries. For example, a test for an addition function might ask what "2+2" evaluates to (4) or if the function can process "ab + cd" as input (it should not unless there is expected behavior for adding strings). These tests report on a pass/fail level, which makes results simple to interpret. However useful, these tests cannot capture interactions between objects. This is the role of integration tests, which in Rosetta are implemented as subsequent comparisons of simulation output for very brief simulations. These tests report whether or not a change is detected, so a developer can track the effects of their code modifications. The tests are short so they can be run every time code is edited. As a consequence, integration tests cannot inform on large scale effects of changes. For example, after updating the RosettaAntibody code, the integration test of homology modeling PDB 1ZTX might change, but that would not give information about RosettaAntibody's performance on antibody modeling in general. To understand how changes affect large-scale performance, the Rosetta community relies on scientific tests.

Table 3.4: Summary of antibody-related scientific benchmarks.

Test	Executable	Metrics	Set
Antibody Grafting	`antibody`	OCD, and RMSD of H1, H2, H3, FRH, L1, L2, L3, FRL	49 antibodies
Antibody CDR-H3 Loop Modeling	`antibody_H3`	CDR-H3 Loop RMSD, Kink	49 grafted models
Antibody–Antigen Docking	`snugdock`	Interface Score, RMSD	15 complexes

Scientific tests typically amount to recreating published results, as they aim to evaluate the performance of a particular modeling protocol (*e.g.* RosettaAntibody) on a comprehensive set of targets (*e.g.* 49 antibodies). For this reason, scientific tests are computationally expensive, but also the best indicator for modeling accuracy. Automated interpretation of scientific results is challenging as one has to quantify what a "good" test outcome might constitute. In the process of making the fundamental alterations to antibody modeling and docking code I described in this chapter, I have also assessed the updated codes' performance on scientific tests, and I am working with Dr. Julia Koehler Leman (of the Flatiron Institute) to standardize these tests within the Rosetta software suite and to setup the tests on the testing servers. The tests are summarized in Table 3.4.

The test for RosettaAntibody consists of 49 antibodies, the selection of which is fully detailed in the methods of two prior papers[8,25]. Briefly, the antibodies were selected for unique CDR-H3 loop sequences and high resolution (better than 2.5 Å) in addition to other quality criteria. The test comprises two stages: grafting and H3 modeling. As of the writing of this thesis, PDB IDs 3NPS and 3MLR are omitted. 3NPS is missing key residues in heavy framework and 3MLR has an extra long CDR-L3 loop with no other possible templates, so an realistic estimation of its grafting accuracy is not feasible.

For the grafting stage, the FASTA sequence of each antibody is used as input and a single homology model is generated, while excluding the input PDB ID as a possible grafting source (see the Appendix). Each model is then compared to the crystal structure. The following structural metrics are compared: the orientation coordinates, calculated as

previously described[8], and RMSDs for structural regions on the light and heavy chain, which are extracted following alignment on conserved framework positions for the respective chains[x].

A sample outcome for the scientific benchmark of grafting is shown in Figure 3.7, which compares the results for grafting before and after the automation of the antibody database. In the case of Figure 3.7, the benchmark was run twice (once with each database) and the RMSDs are compared. When implemented as a server-based, stand-alone test, the benchmark will not be able to compare against itself. Instead, quantifiable metrics, such as the fraction sub-Ångström models for each region will report a pass or fail status based on a threshold. For each of the framework regions, I would set a threshold at one model with RMSD greater than 1 Å. I would exclude the CDR-H3 loop from this analysis as the goal of grafting is not to yield an accurate CDR-H3 loop model. For the remaining CDRs, I would anticipate no more than 5 of the models (for each CDR) to have RMSD higher than 1 Å.

Grafted models are used as input for the H3 modeling test (because the ultimate challenge for RosettaAntibody is to produce an accurate model from sequence alone, it is nonsensical to test H3 modeling on a crystal structure). For each input model, 1000 CDR-H3 loop models are generated with RosettaAntibody's default H3 modeling approach (the command line titled "Antibody H3 Kink Constraints" in the Appendix). Following modeling, CDR-H3 loop RMSDs are calculated by aligning the heavy chain of the model to the crystal structure and comparing the positions of the heavy atoms.

The result for the CDR-H3 loop modeling benchmark were shown earlier in this chapter in Figure 3.4. Here, the benchmark was run twice. First, with the standard next-generation KIC method for loop closure, then with the new fragment-based method. In Figure 3.4, I compared the minimum RMSD observed across all models for each antibody target, as I sought to answer whether or not fragments improved CDR-H3 Loop modeling (they did not). When implemented on the testing servers, the benchmark will not run as a comparison.

[x]To ensure accurate comparison, the light and heavy are not aligned simultaneously, but rather independently.

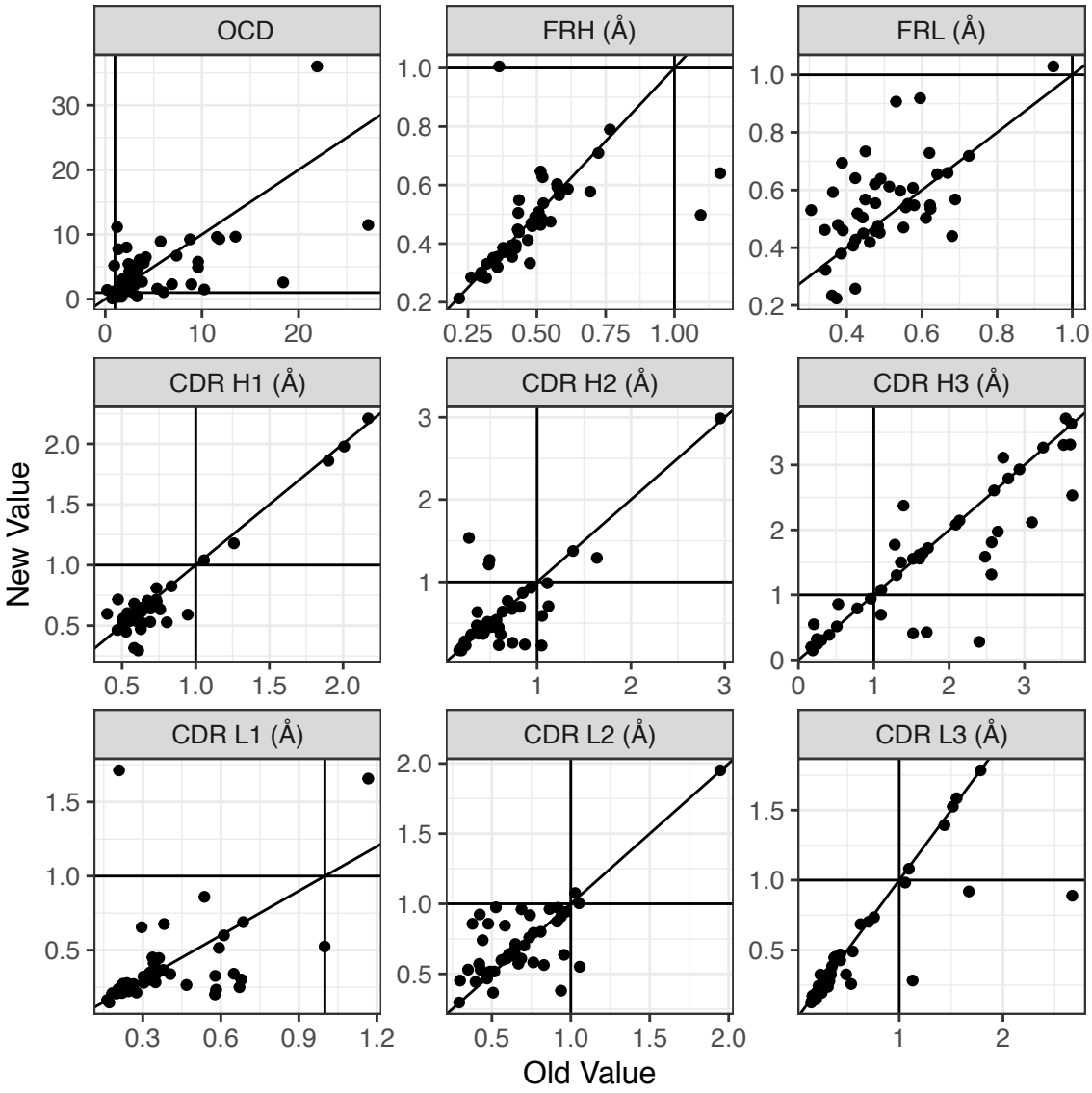

Figure 3.7: Values for structural metrics with respect to native following the grafting step with either the old (manual, x-axis) or new (automatic, y-axis) database are plotted. The new database slightly improves the performance of the grafting step of RosettaAntibody, with 55% of CDR loops and 53.5% of FRs having lower RMSDs.

Instead, the minimum CDR-H3 loop RMSD values for each targets will be binned and the number of sub-X Ångström models will be compared to a standard. For example, I expect between zero and one loops to have an RMSD of > 3.0 Å, between five and ten loops to have an RMSD of $3.0 \leq X > 2.0$ Å, between 15 and 20 loops to have an RMSD of $2.0 \leq X > 1.0$ Å, and between 10 and 15 loops to have an RMSD of $1.0 \leq X > 0.0$ Å. As the CDR-H3 loop modeling is stochastic in nature, these values are estimates and will need to be calibrated over time.

The test for Rosetta SnugDock consists of 15 antibody–antigen complexes[xi], for some of which unbound structures are available, established in the original SnugDock paper[18]. The input for the SnugDock scientific test is the ten antibody homology models produced by the default RosettaAntibody protocol[26], while excluding the native PDB from grafting, and a backbone-constrained relaxed[27] (unbound when possible) antigen structure. The input structures are prepacked[3] in preparation for docking, using the `docking_prepack_protocol` application (sample commands are shown in the Appendix). Finally, SnugDock is run, when possible using Motif Dock Score to improve low-resolution sampling[28]. Afterwards, models are aligned to crystal structures and the interface RMSD is computed, along with the interface score.

A sample result for the SnugDock scientific benchmark is shown in Figure 3.8. This figure exemplifies the challenges of assessing scientific benchmarking results (an issue that exists for the CDR-H3 loop modeling benchmark as well, but was not discussed). By eye, the user can see good performance (low-scoring, low-RMSD models with few/no false positives [low-scoring, but high-RMSD models]) of the docking algorithm for 5 targets (1ahw, 1jps, 1mlc, 1ynt, and 1ztx), with decent performance for a further 5 targets (1bql, 2aep, 2b2x, 2bdn, and 2jel), and poor performance for the remaining 5 targets (1jhl, 1k4c, 1nca, 1wej, and 1vfb). However, there is no single quantifiable metric that can capture comprehensively capture these observations.

[xi] PDB IDs: 1AHW, 1BQL, 1JHL, 1JPS. 1K4C, 1MLC, 1NCA, 1VFB, 1WEJ, 1YNT, 1ZTX, 2AEP, 2B2X, 2BDN, and 2JEL

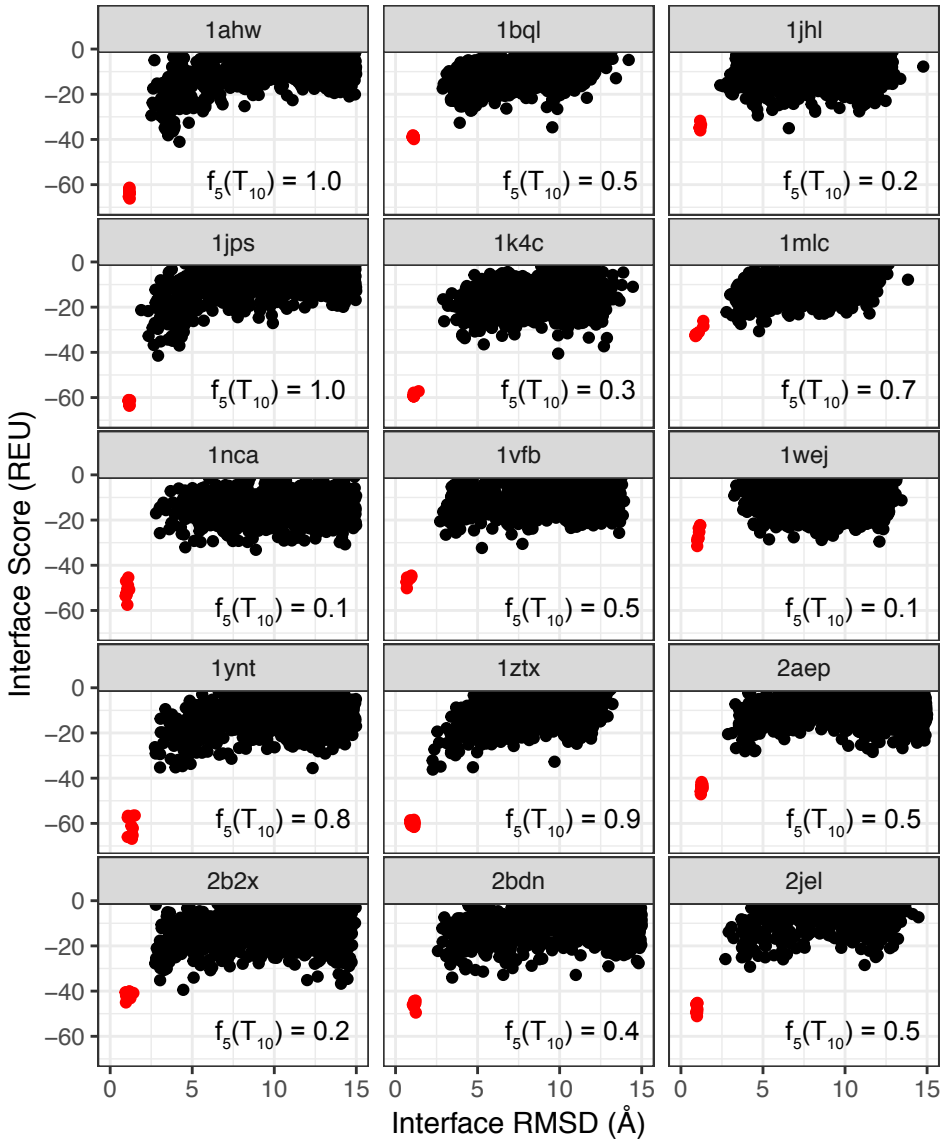

Figure 3.8: "Funnel plots" show interface RMSD versus interface score for models produced by docking simulations (here 15 antibody–antigen complexes, with PDB IDs specified for each sub-plot). If the modeling protocol is properly calibrated, low score will trend with low RMSD and hence points will "funnel" towards the bottom left of each sub-plot. Native structures are scored as well (red points), with minor refinement, as a control. "Funnel plots" address two questions about docking: (1) does the simulation sample native-like (low-RMSD) conformations, for example 2bdn does not, and (2) does the simulation score these points accurately, for example 1wej does not. The value in the bottom right is the fraction of top 10 lowest-scoring models with an interface RMSD less than 5 Å.

As with the CDR-H3 benchmark, one could use minimum-RMSD values as a proxy for sampling, answering whether or not the simulations are exploring near-native conformations. But such an approach would not be comprehensive. It contains no information about performance in blind cases, ones where the user would select low-scoring models, assuming these models were representative of the native structures. Therefore, an alternative approach is to quantify the number of models below a certain RMSD from the top X lowest-scoring, then set a threshold for success. Based on the CAPRI criteria for a "medium" quality complex models, I propose the following definition of success: 50% of the top 10 lowest-scoring models have an interface RMSD of less than 5 Ångströms (this value is shown in the bottom of each sub-plot of Figure 3.8).

3.8 Summary

In this chapter, I reported the recent technical advances I have contributed to RosettaAntibody and Rosetta SnugDock. To encourage future development, I made the code more stable by converting old Python scripts to object-oriented C++ classes, I expanded the homology database and enabled its automatic updating, and I curated scientific benchmarks. To tackle new classes of antibodies (such as those with only a single heavy chain), I made RosettaAntibody and Rosetta SnugDock more robust, implemented new loop modeling methods, and developed a Hierarchical FoldTree. Finally, I eased the use of future users by simplifying the user interface. Collectively, these changes enabled multiple future studies[4], including the one reported in Chapter 4.

3.A Appendix

3.A.1 Supplemental figures

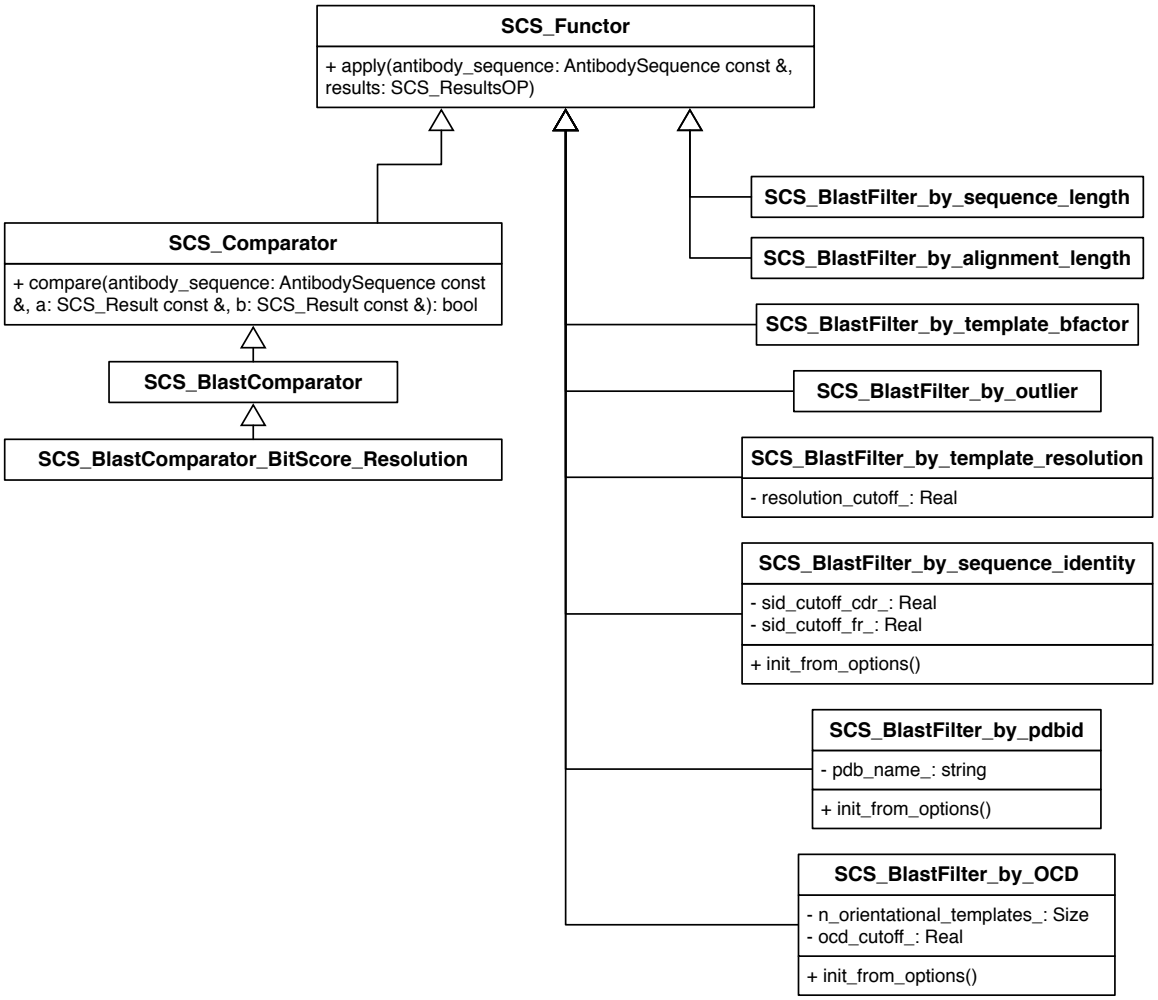

Figure 3.A.1: UML diagram of the SCS_Functor and associated classes for filtering antibody templates. In a UML diagram, classes are represented by boxes. Data methods (of the form `attribute: type`) and members (of the form `function(args): return`) follow in subsequent boxes. The first character indicates the visibility ("+": public and "-": private). Open triangle connections indicate inheritance.

Figure 3.A.2: UML diagram of the SCS_ResultSet and associated classes for storing potential template data. In a UML diagram, classes are represented by boxes. Data methods (of the form `attribute : type`) and members (of the form `function(args): return`) follow in subsequent boxes. The first character indicates the visibility ("+": public). Open triangle connections indicate inheritance and diamonds indicate membership. Note: diamonds should be flipped.

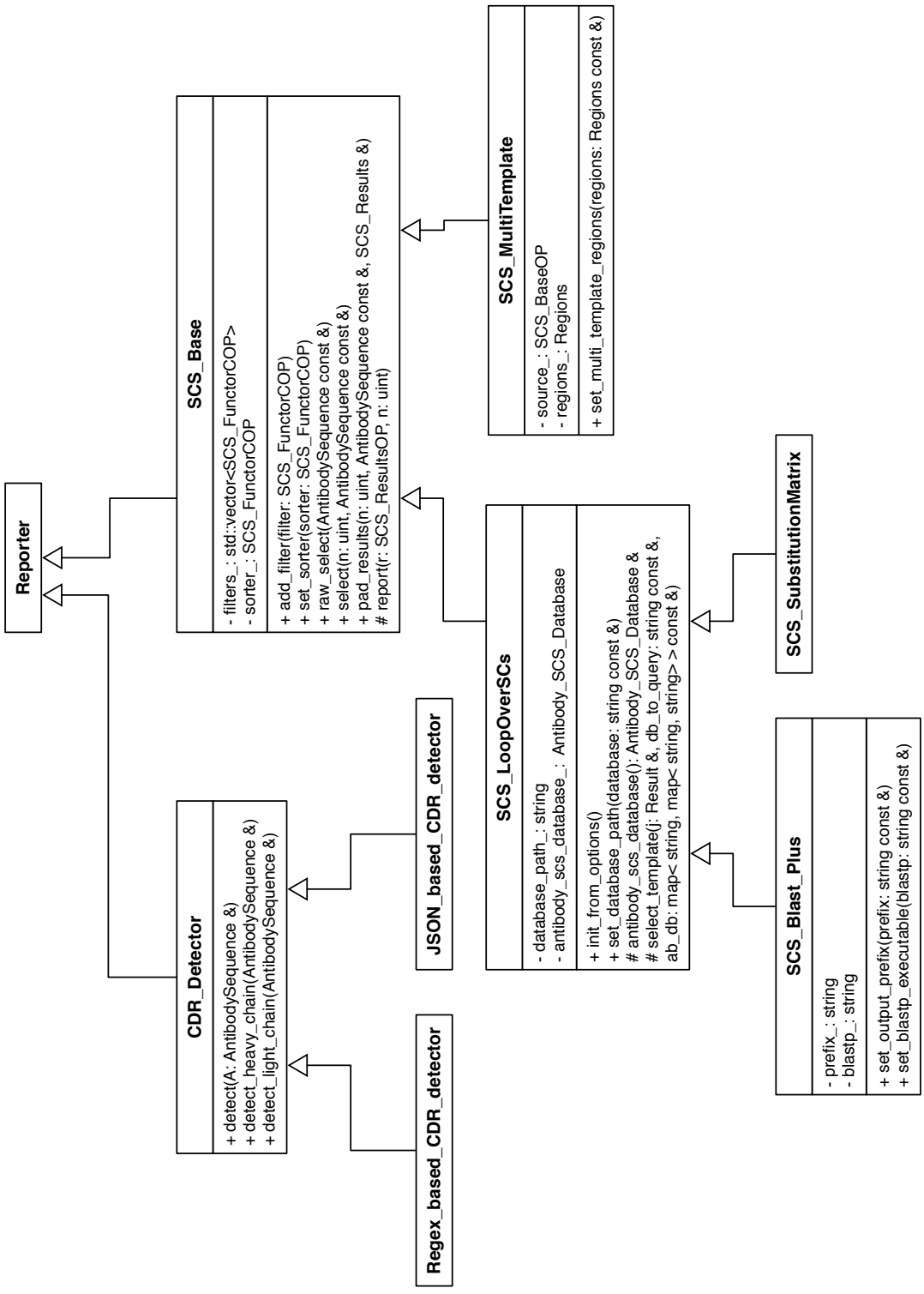

Figure 3.A.3: UML diagram of the Reporter-derived classes. The CDR_Detector identifies CDRs from sequences based on rules defined by the inheriting subclass. The SCS_Base class defines the necessary virtual function for selecting structural templates. In a UML diagram, classes are represented by boxes. Data methods (of the form `attribute : type`) and members (of the form `function(args): return`) follow in subsequent boxes. The first character indicates the visibility ("+": public, "-": private, "#":protected). Open triangle connections indicate inheritance.

3.A.2 Sample Commands to Run RosettaAntibody and SnugDock

These commands are valid for the git branch: `lqtza/enable_fkic_in_antibody_H3` (9f10be3). Once merged with this branch is merged with the master branch, the commands will be valid for the public release of Rosetta. Note that these commands do not include the Q–Q constraint because it has to be manually specified via a constraint file. The Q–Q constraint should be automated in the future.

Antibody Grafting

```
antibody.linuxgccrelease -fasta my.fasta
```

Antibody Grafting Exclude PDB

```
antibody.linuxgccrelease -fasta my.fasta -antibody:exclude_pdb PDBID
```

Antibody H3 Minimal

```
antibody_H3.linuxgccrelease -s model.relaxed.pdb -nstruct 1000
```

Antibody H3 Extra Rotamers

```
antibody_H3.linuxgccrelease
  -s model.relaxed.pdb
  -nstruct 1000
  -ex1
  -ex2
  -extrachi_cutoff 0
```

Antibody H3 Kink Constraints

```
antibody_H3.linuxgccrelease
  -s model.relaxed.pdb
  -nstruct 1000
  -antibody:constrain_cter
  -antibody:auto_generate_kink_constraint
  -antibody:all_atom_mode_kink_constraint
  -ex1
  -ex2
  -extrachi_cutoff 0
```

Antibody H3 Fragment KIC

```
antibody_H3.linuxgccrelease
  -s model.relaxed.pdb
```

```
  -nstruct 1000
  -antibody:constrain_cter
  -antibody:auto_generate_kink_constraint
  -antibody:all_atom_mode_kink_constraint
  -ex1
  -ex2
  -extrachi_cutoff 0
  -loops:frag_sizes 9 3 1
  -loops:frag_files 9mers 3mers none
```

SnugDock Minimal

```
snugdock.linuxgccrelease
  -s input.pdb
  -partners G_HL
  -spin
  -dock_pert 3 8
  -detect_disulf false
  -nstruct 1000
```

SnugDock Minimal With Kink Constraint

```
snugdock.linuxgccrelease
  -s input.pdb
  -partners G_HL
  -spin
  -dock_pert 3 8
  -detect_disulf false
  -antibody:auto_generate_kink_constraint
  -antibody:all_atom_mode_kink_constraint
  -nstruct 1000
```

SnugDock Extra Rotamers

```
snugdock.linuxgccrelease
  -s input.pdb
  -partners G_HL
  -spin
  -dock_pert 3 8
  -detect_disulf false
  -ex1
  -ex2aro
  -nstruct 1000
```

SnugDock Ensemble

```
snugdock.linuxgccrelease
  -s input.pdb
  -ensemble1 ag.list
  -ensemble2 h3.list
```

```
  -partners G_HL
  -spin
  -dock_pert 3 8
  -detect_disulf false
  -ex1
  -ex2aro
  -nstruct 1000
```

SnugDock Motif Scores

```
snugdock.linuxgccrelease
  -s input.pdb
  -ensemble1 ag.list
  -ensemble2 h3.list
  -partners G_HL
  -spin
  -dock_pert 3 8
  -detect_disulf false
  -ex1
  -ex2aro
  -docking_low_res_score motif_dock_score
  -mh:path:scores_BB_BB motif_dock/score_data_
  -mh:score:use_ss1 false
  -mh:score:use_ss2 false
  -mh:score:use_aa1 true
  -mh:score:use_aa2 true
  -nstruct 1000
```

3.A.3 Auxillary Commands to Run SnugDock/Prepack

Relax

```
relax.linuxgccrelease
  -s antigen.pdb
  -ex1
  -ex2
  -use_input_sc
  -relax:constrain_relax_to_start_coords
  -relax:ramp_constraints false
```

Prepack

```
docking_prepack_protocol.linuxgccrelease
  -s input.pdb
  -docking:partners G_HL
  -docking::dock_rtmin
  -docking::sc_min
  -ensemble1 ag.list
  -ensemble2 h3.list
```

Docking Local Refine

```
docking_protocol.macosclangrelease
  -s input.pdb
  -docking:docking_local_refine
  -ex1
  -ex2aro
  -partners A_B
  -nstruct 10
```

3.A.4 Antibody Modeling Benchmark List

```
1dlf,1fns,1gig,1jfq,1jpt,1mfa,1mlb,1mqk,1nlb,1oaq,1seq,1x9q,2adf,
2d7t,2e27,2fb4,2fbj,2r8s,2v17,2vxv,2w60,2xwt,2ypv,3e8u,3eo9,3g5y,
3giz,3gnm,3go1,3hc4,3hnt,3i9g,3ifl,3liz,3lmj,3m8o,3mlr,3mxw,3nps,
3oz9,3p0y,3t65,3umt,3v0w,4f57,4h0h,4h20,4hpy,4nzu
```

References

1. Weitzner, B. D., Kuroda, D., Marze, N., Xu, J. & Gray, J. J. Blind prediction performance of RosettaAntibody 3.0: Grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins: Structure, Function and Bioinformatics* **82,** 1611–1623 (2014).
2. Stein, A. & Kortemme, T. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS ONE* **8** (ed Zhang, Y.) e63090 (2013).
3. Chaudhury, S. & Gray, J. J. Conformer Selection and Induced Fit in Flexible Backbone Protein–Protein Docking Using Computational and NMR Ensembles. *Journal of Molecular Biology* **381,** 1068–1087 (2008).
4. Høydahl, L. S. *et al.* Plasma Cells are the Most Abundant Gluten Peptide MHC-expressing Cells in Inflamed Intestinal Tissues From Patients With Celiac Disease. *Gastroenterology* (2018).
5. DeKosky, B. J. *et al.* Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proceedings of the National Academy of Sciences* **113,** E2636–E2645 (2016).
6. Kovaltsuk, A. *et al.* How B-cell receptor repertoire sequencing can be enriched with structural antibody data. *Frontiers in Immunology* **8,** 1753 (2017).
7. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25,** 3389–402 (1997).
8. Marze, N. A., Lyskov, S. & Gray, J. J. Improved prediction of antibody VL-VH orientation. *Protein Engineering, Design and Selection* **29,** 409–418 (2016).
9. Adolf-Bryfogle, J. *et al.* RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *PLoS Computational Biology* **14** (ed Ben-Tal, N.) e1006112 (2018).
10. Sivasubramanian, A., Chao, G., Pressler, H. M., Wittrup, K. D. & Gray, J. J. Structural model of the mAb 806-EGFR complex using computational docking followed by computational and experimental mutagenesis. *Structure* **14,** 401–414 (2006).
11. Dunbar, J. *et al.* SAbDab: The structural antibody database. *Nucleic Acids Research* **42,** D1140–D1146 (2014).
12. Chaudhury, S., Lyskov, S. & Gray, J. J. *PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta* 2010.

13. Abhinandan, K. R. & Martin, A. C. R. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Molecular Immunology* **45,** 3832–3839 (2008).

14. Sircar, A., Sanni, K. A., Shi, J. & Gray, J. J. Analysis and Modeling of the Variable Region of Camelid Single-Domain Antibodies. *The Journal of Immunology* **186** (2011).

15. Leaver-Fay, A. *et al.* Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology. Computer Methods, Part C* **487** (eds Brand, M. L. J. & Ludwig) 545–574 (2011).

16. Wang, C., Bradley, P. & Baker, D. ProteinâĂŞProtein Docking with Backbone Flexibility. *Journal of Molecular Biology* **373,** 503–519 (2007).

17. Parsons, J., Bradley Holmes, J., Maurice Rojas, J., Tsai, J. & Strauss, C. E. M. Practical conversion from torsion space to cartesian space for in silico protein synthesis. en. *Journal of Computational Chemistry* **26,** 1063–1068 (2005).

18. Sircar, A. & Gray, J. J. SnugDock: Paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Computational Biology* **6,** e1000644 (2010).

19. Marze, N. A. *Building Computational Tools for Antibody Modeling and Protein–Protein Docking* PhD thesis (Johns Hopkins University, 2017), 133.

20. Almagro, J. C. *et al.* Second antibody modeling assessment (AMA-II). *Proteins: Structure, Function, and Bioinformatics* **82,** 1553–1562 (2014).

21. Collins, A. M. & Jackson, K. J. *On being the right size: antibody repertoire formation in the mouse and human* 2018.

22. Regep, C., Georges, G., Shi, J., Popovic, B. & Deane, C. M. The H3 loop of antibodies shows unique structural characteristics. *Proteins: Structure, Function and Bioinformatics* **85,** 1311–1318 (2017).

23. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods* **6,** 551–552 (2009).

24. Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. & Baker, D. Generalized fragment picking in rosetta: Design, protocols and applications. *PLoS ONE* **6** (ed Uversky, V. N.) e23294 (2011).

25. Weitzner, B. D. & Gray, J. J. Accurate Structure Prediction of CDR H3 Loops Enabled by a Novel Structure-Based C-Terminal Constraint. *The Journal of Immunology* **198,** 505–515 (2016).

26. Weitzner, B. D. *et al.* Modeling and docking of antibody structures with Rosetta. *Nature Protocols* **12,** 401–416 (2017).

27. Nivon, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS ONE* **8** (ed Zhang, Y.) e59004 (2013).

28. Marze, N. A., Roy Burman, S. S., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics* **34** (ed Valencia, A.) 3461–3469 (2018).

Chapter 4

Large-Scale Antibody CDR-H3 Loop Flexibility Assessment

4.1 Overview

Antibodies can rapidly evolve in specific response to antigens. Affinity maturation drives this evolution through cycles of mutation and selection leading to enhanced antibody specificity and affinity. Elucidating the biophysical mechanisms that underlie affinity maturation is fundamental to understanding B-cell immunity. An emergent hypothesis is that affinity maturation reduces the conformational flexibility of the antibody's antigen-binding paratope to minimize entropic losses incurred upon binding. In recent years, computational and experimental approaches have tested this hypothesis on a small number of antibodies, often observing a decrease in the flexibility of the Complementarity Determining Region (CDR) loops that typically comprise the paratope and in particular the CDR-H3 loop, which contributes a plurality of antigen contacts. However, there were a few exceptions, and

previous studies were limited to a small handful of cases. I[i] determined the structural flexibility of the CDR-H3 loop for thousands of recently-determined homology models of the human peripheral blood cell antibody repertoire using rigidity theory. I found no clear delineation in the flexibility of naïve and antigen-experienced antibodies. To account for possible sources of error, I additionally analyzed hundreds of human and mouse antibodies in the Protein Data Bank through both rigidity theory and B-factor analysis. By both metrics, I observed only a slight decrease in the CDR-H3 loop flexibility when comparing affinity-matured antibodies o naïve antibodies, and the decrease was not as drastic as previously reported. Further analysis, incorporating molecular dynamics (MD) simulations, revealed a spectrum of changes in flexibility. My results suggest that rigidification may be just one of many biophysical mechanisms for increasing affinity.

4.2 Introduction

Antibodies are proteins produced by the B cells of jawed vertebrates that play a central role in the adaptive immune system. They recognize a variety of pathogens and induce further immune response to protect the organism from external perturbation. Molecules that are bound by antibodies are referred to as antigen and are recognized by the antibody variable domain (Fv), which is comprised of a variable heavy (VH) and light (VL) domain. To overcome the challenge of recognizing a vast array of targets — the number of antigens being far greater than the number of antibody germline genes — antibodies rely on combinatoric and genetic mechanisms that increase sequence diversity[1–3]. Starting from a limited array of germline genes, a naïve antibody is generated by productive pairing of a randomly recombined VH, assembled from V-, D-, and J-genes on the heavy locus, and randomly recombined VL, assembled from V- and J-genes on the kappa and lambda loci[1]. Next, in a process known as affinity maturation, iterations of somatic hypermutation are followed by selection to evolve the antibody in specific response to a particular antigen. This

[i] While I guided the research in general, worked on antibody modeling, and analyzed results, Adnan Sljoka developed the graph theoretical approach, and Daisuke Kuroda ran the molecular dynamics simulations.

evolution results in the gradual accumulation of mutations across the entire antibody, with higher mutation rates in the six complementarity determining regions (CDRs) than in the framework regions (FRs)[4,5]. The CDRs are hypervariable loops comprising a binding interface on the Fv domain beta-sandwich framework, with three loops contributed by each chain; the light chain CDRs are denoted as L1, L2, and L3 and the heavy chain CDRs are H1, H2, and H3. The five non-H3 CDRs can be readily classified into a discrete amount of canonical structures[6–10] because they possess limited diversity in both sequence and structure. The CDR-H3 on the other hand is the focal point of V(D)J recombination, resulting in exceptional diversity of both structure and sequence. While all CDRs contribute to antigen binding, the diverse CDR-H3 is often the most important CDR for antigen recognition[11–14]. Thus, to understand the role of B cells in adaptive immunity and how they evolve antibodies capable of binding specific antigens, we must first understand the effects of affinity maturation on the CDRs, and in particular on the CDR-H3.

Over the last 20 years, the effects of affinity maturation have been studied with an assortment of experimental and computational methods. X-ray crystallography has been used to compare antigen-inexperienced (naïve) and antigen-experienced (mature) antibodies with both antigen present and absent. Analysis of the catalytic antibodies 48G7, AZ-28, 28B4, and 7G12 showed a 1.2 Å average increase in Cα RMSD of the CDR-H3 upon antigen binding in the naïve over that of the mature antibody, whereas motion in the other CDRs varied[15–19]. Beyond structural studies, surface plasmon resonance (SPR) has been used to assess the energetics and association/dissociation rate constants of antibody–antigen binding. Manivel *et al.* studied a panel of 14 primary (naïve) and 11 secondary (mature) response anti-peptide antibodies, observing that affinity maturation resulted in increases in the association rate and corresponding changes in the entropy of binding[20]. Schmidt *et al.* saw the opposite when studying a broadly neutralizing influenza virus antibody, observing that affinity maturation resulted primarily in a decrease in the dissociation rate, with little effect on the association rate[21]. Isothermal calorimetry (ITC) has also been used to determine

antigen-binding energetics including the enthalpic and entropic contributions. For nine anti-fluorescein antibodies, including 4-4-20 and eight anti-MPTS antibodies, ITC results revealed diverse effects of affinity maturation: 14 of 17 mature antibodies bound antigen in an enthalpically favorable and entropically unfavorable manner, yet 3 of 17 showed the opposite, with entropically favorable and enthalpically unfavorable binding energetics[22,23]. Three-pulse photon echo peak shift (3PEPS) spectroscopy has been used to quantify dynamics of chromophore-bound antibodies on short timescales of femto- to nanoseconds. 3PEPS spectroscopy results from a panel of 18 antibodies showed that mature antibodies can possess a range of motions from small rearrangements such as side-chain motions to large rearrangements such as loop motions[22–24]. In a specific comparison of naïve vs. mature, for the 4-4-20 antibody, the mature antibody was found to have smaller motions, i.e. to be more rigid, than naïve[22–27]. Antibody dynamics have also been studied by hydrogen–deuterium exchange mass spectroscopy (HDX-MS), which in contrast to 3PEPS probes timescales of seconds to hours. Comparison of three naïve and mature anti-HIV antibodies showed changes in CDR-L2/H2, but not in CDR-H3 dynamics[28]. Finally, MD simulations have been used to study antibody dynamics on intermediate timescales of nano- to microseconds. MD simulations showed rigidification and reduction of CDR-H3 loop motion upon maturation for seven studied naïve/mature antibodies, with two exceptions, depending on the specific study[21,27,29–33] (22, 28, 30-34). In an orthogonal protein design approach to examine the CDR-H3 loop flexibility, Babor *et al.* and Willis *et al.* found that naïve antibody structures are more optimal for their sequences, when considering multiple CDR-H3 loop conformations[34,35]. In sum, past studies focusing on the effects of affinity maturation on CDRs have found evidence suggesting that mature antibodies have more structural rigidity and less conformational diversity than their naïve counterparts[15,17,18,22–26].

With recent growth in the number of antibody structures deposited in the Protein Data Bank (PDB) and development of homology models from high-throughput sequencing of paired VH–VL genes in B cells, we now have the datasets necessary to test the rigidity

hypothesis on a large scale. Prior studies, usually focused on a few antibodies at time, generally support the hypothesis that affinity maturation rigidifies the CDR-H3 loop. Thus, I hypothesized that this effect should be observable in a repertoire-scale study of thousands of antibodies. I first analyzed thousands of recently-determined RosettaAntibody homology models of the most common antibody sequences found in the human peripheral blood cell repertoire[36]. I estimated the structural flexibility of the CDR-H3 loop by applying graph theoretical techniques based on mathematical rigidity theory, namely the Floppy Inclusions and Rigid Substructure Topography (FIRST) and extensions of the Pebble Game (PG) algorithms to determine backbone degrees of freedom (DOFs). Surprisingly, I found no difference in the CDR-H3 loop flexibility of the naïve and mature antibody repertoires. I considered alternative explanations for my results, which were incongruent with past studies, by expanding my analysis to a large set of antibody crystal structures, including several previously characterized antibodies, and extending my methods to include other measures of flexibility such as B-factors and MD simulations. By all analysis methods, I found mixed results: some antibodies' CDR-H3 loops were more flexible after affinity maturation whereas others' became less flexible. In summary, I find that while affinity maturation can modulate antibody binding activity by reducing CDR-H3 structural flexibility, it does not necessarily have to do so.

4.3 Methods

4.3.1 Immunomic repertoire modeling

Briefly, RosettaAntibody is an antibody modeling approach that assembles homologous structural regions into a rough model and then refines the model through gradient-based energy minimization, side-chain repacking, rigid-body docking, and *de novo* loop modeling of the CDR-H3. The approach is fully detailed in other publications[37,38]. In a typical simulation, 1,000 models are generated and the ten lowest-energy models are retained. The immunomic repertoire I analyzed is from DeKosky and Lungu, *et al.*[36]. In that study,

models were generated for each of the 1,000 most frequently occurring naïve and mature antibody sequences from two donors (a total of 20,000 models representing the 2,000 most frequent antibodies).

4.3.2 Structural rigidity determination

The flexibility or rigidity of the CDR-H3 loop backbone was determined by using several extensions of the Pebble Game (PG) algorithm[39–42] and method FIRST[43]; I refer to here as FIRST-PG. This approach can determine flexible and rigid regions in a protein and quantify the internal conformational degrees of freedom from a single protein conformational snapshot. FIRST generates a molecular constraint network (*i.e.* a graph) consisting of vertices (nodes) representing atoms and edges (interactions representing covalent bonds, hydrogen bonds, hydrophobic interactions, etc.). Each potential hydrogen bond is assigned an energy in kcal/mol which is dependent on donor-hydrogen–acceptor geometry. FIRST is run with a selected hydrogen-bonding energy cutoff, where all bonds weaker than this cutoff are ignored in the network. On the resulting network, the well-developed mathematical and structural engineering concepts[44] of flexibility and rigidity of molecular frameworks and the PG algorithm are then used to identify rigid clusters, flexible regions, and overall available conformational DOFs. For a given antibody structure, DOFs for the protein backbone of the CDR-H3 loop were calculated at every hydrogen-bonding energy cutoff value between 0 to -7 kcal/mol in increment steps of 0.01 kcal/mol. This calculation was repeated for every member of that antibody ensemble (*i.e.* ten lowest energy models of the ensemble) and finally, at each energy cutoff, the DOF count was averaged over the entire ensemble.

For a given energy cutoff and a given member of the ensemble, the DOF count for the CDR-H3 loop (residues 95–102) was obtained using a special PG operation which calculates the maximum number of pebbles that can be gathered on the backbone atoms (Cα, C, N) of the CDR-H3 loop[39]. The PG algorithm starts with the constrained molecular graph and generates a directed multigraph, where available free pebbles are absorbed one by one by

independent edges (constraints). Each pebble represents one of 6 DOF associated with an atom. After PG completion, the remaining free pebbles that can be collected on the CDR-H3 backbone (*i.e.* a subgraph in the constrained network) represent its conformational DOF count.

4.3.3 Degree of freedom scaling

To compare flexibility across CDR-H3 loops of different lengths, the DOF metric computed above is scaled by a theoretical maximum DOF. I define $sDOF = \frac{DOF}{(2L+6)}$, where, $2L$ (the loop length in residues) represents the backbone degrees of freedom (torsion angles: ϕ,ψ), and 6 represents the trivial but ever-present rigid-body DOFs (*i.e.* combination of rotations and translations in 3D).

4.3.4 Area under the curve calculation

The area under the curve (AUC) is approximated by simple numerical integral (akin to trapezoidal integration), where the first term defines a rectangle and the second term defines a triangle:

$$AUC = \sum (x_i - x_{i-1}) \cdot y_{i-1} + \frac{1}{2}(x_i - x_{i-1})(y_i - y_{i-1})$$

4.3.5 Crystallographic dataset

On June 27th, 2017, a summary file was generated from the Structural Antibody Database (SAbDab)[45], using the "non-redundant search" option to search for antibodies with maximum 99% sequence identity, paired heavy and light chains, and a resolution cutoff of 3.0 Å. The summary file, containing 1021 antibodies, was used as input to a SAbDab download script which yielded corresponding sequences, Chothia-numbered PDBs, and IMGT data (on occasion this had to be updated to match the reported germline in the IMGT 3Dstructure-DB)[46]. The structures were further pruned: structures were omitted if there

were unresolved CDR-H3 residues, as this would preclude flexibility calculations, or if the antibody was neither human nor mouse, as this would prevent alignment to germline. Prior to analysis, structures were truncated to the Fv region (removing all residues but light chain residues numbered 1–108 and heavy chain residues numbered 1–112, in Chothia numbering) and duplicate and non-antibody (for example, bound antigen) chains were removed. A total of 922 antibody crystal structures were analyzed. The following CDR definitions were used throughout this paper, in conjunction with the Chothia numbering scheme: L1 spans light chain residue numbers 24–34, L2 spans 50–56, L3 spans 89–97, H1 spans heavy chain residue numbers 26–35, H2 spans 50–56, and H3 spans 95–102.

4.3.6 Alignment to germline

The germline of each antibody was determined by IMGT lookup[46]. Then, BLASTP (version 2.2.29+) with the BLOSUM50 scoring matrix was used to align the antibody variable region heavy and light sequences to corresponding germline sequences (IGHV, IGKV, and IGLV loci only, downloaded from IMGT). The number of mismatches according to BLAST were considered as the number of amino acid mutations from germline. Supplementary Table 1 in the original publication[47] details the PDB ID, CDR-H3 length, number of heavy chain mutations, number of light chain mutations, heavy germline gene, and light germline gene data for each structure in the dataset.

4.3.7 B-factor Z-score calculation

Temperature factors (B-factors) were extracted for all Cα atoms in the variable region of the antibody heavy chain (VH, Chothia numbering 1–112). The arithmetic mean and sample standard deviation values were calculated for the B-factors. For each Cα atom in the CDR-H3 region, residue numbers spanning 95–102 under the Chothia numbering convention[11], the z-score was calculated as $(x - \mu)/\sigma$, where x is the B-factor of the current Cα atom and μ and σ are the mean and standard deviation of B-factors for all Cα atoms in the VH,

respectively. PDB IDs 2NR6 and 3HAE were excluded from B-factor analysis because all reported B-factors were identical and so the z-scores were zero by definition.

To test whether two observed B-factor distributions arose from the same underlying distribution, I turned to randomization testing. First, I computed the difference of the observed distribution means. Next, I pooled the data from the two distributions (e.g. CDR-H3 loop B-factor z-scores) and randomly sampled the pooled data to create two simulated distributions (e.g. randomly assigning z-scores to either the naïve or mature category). Finally, I computed the simulated difference of the randomized distribution means. This process was repeated 10,000 times, so that I could identify the fraction of random distributions with differences greater than the observed. Since this process is stochastic and does not exhaustively sample all permutations of the data, it was further repeated 10 times to acquire a standard deviation.

4.3.8 Rosetta relaxation and ensemble generation

Antibody structural ensembles with 10 members were generated using either the Rosetta FastRelax[48] or Rosetta KIC protocol[49], and Rosetta version 2017.26-dev59567 was used for all simulations (corresponding to weekly release version 2017.26). The Rosetta FastRelax protocol consists of five cycles of side-chain repacking and gradient-based energy minimization in the REF2015 version of the Rosetta energy function[50]. Thus, FastRelax ensembles explore the local energy minimum of the crystal structure. KIC ensembles are more diverse and representative of RosettaAntibody homology models: each ensemble member was generated by running the CDR-H3 refinement step of the RosettaAntibody protocol, consisting of VH–VL docking, CDR-H3 loop remodeling, and all-CDR loop minimization[37,38]. Sample command lines are given in the Supplementary Material. The structural ensembles produced by both FastRelax and KIC were used for rigidity analysis. For technical reasons, six targets could not be analyzed from the FastRelax ensemble, and 177 targets from the KIC ensemble were omitted due to non-trivial incompatibilities between the input structure

numbering and Rosetta's internal antibody numbering scheme and a computing cluster time limitation. The excluded targets were randomly distributed and likely would not affect the conclusions.

4.3.9 Molecular dynamics simulations

The Fv regions were retrieved from the original PDB files. The MD simulations were performed using the NAMD 2.12 package[51] with the CHARMM36m force field and the CMAP backbone energy correction[52]. The truncated Fv structures were solvated with TIP3P water in a rectangular box such that the minimum distance to the edge of the box was 12 Å under periodic boundary conditions. Na or Cl ions were added to neutralize the protein charge, then further ions were added corresponding to a salt solution of concentration 0.14 M. The time step was set to 2 fs throughout the simulations. A cutoff distance of 10 Å for Coulomb and van der Waals interactions was used. Long-range electrostatics were evaluated through the Particle Mesh Ewald method[53].

The initial structures were energy-minimized by the conjugate gradient method (10,000 steps), and heated from 50K to 300K during 100 ps, and the simulations were continued by 1 ns with NVT ensemble, where protein atoms were initially held fixed whereas non-protein atoms freely moved, gradually releasing the whole system to facilitate a stable simulation over the 1 ns. Further simulations were performed with NPT ensemble at 300K for 200 ns without any restraints other than the SHAKE algorithm to constrain bonds involving hydrogen atoms. The last 180 ns of each trajectory was used for the subsequent clustering analyses. Similar to a previous work[54], a total of 2000 evenly spaced frames from each trajectory were clustered based on root-mean-square deviation (RMSD) of the Cα and Cβ atoms using the K-means clustering algorithm implemented in the KCLUST module in the MMTSB tool set[55]. The cluster radius was adjusted to maintain 20 clusters in each trajectory. The structure closest to the center of each cluster was chosen as a representative structure of each cluster. The 10 representative structures were chosen from the top 10

largest clusters and these representative structures were energy-minimized by the conjugate gradient method (10,000 steps) in a rectangular water box. The minimized antibody Fv structures were used as the inputs for the rigidity analysis.

Root-mean-square quantities of the MD trajectories were calculated based on the last 180 ns trajectories. After superposing Cα atoms of the FR of the heavy chain (FRH) of each snapshot onto Cα atoms of FRH of the reference structures (i.e. crystal structures), Cα-RMSD of the CDR-H3 loop was calculated as the time average. Similarly, after superposing Cα atoms of entire Fv domains of each snapshot onto those of the reference structures, the root-mean-square fluctuation (RMSF) of a residue i was defined as the time average:

$$RMSF_i = \sqrt{\langle (x_i - \langle x_i \rangle)^2 \rangle}$$

where x_i is the distance between the Cα atom of the snapshots at a given time and the Cα atom of the ith residue of the reference structures[56].

4.4 Results

4.4.1 Immunomic repertoire reveals no difference in flexibility between naïve and mature CDR-H3 loops

I initially asked whether CDR-H3 loop rigidification, having been observed in many past studies, was present in a large set of antibodies derived from human peripheral blood cells. Previously, DeKosky and Lungu *et al.* used RosettaAntibody to model the structures of 1,911 common antibodies found in the peripheral blood cells of two human donors[36]. Paired VH–VL sequences were derived from either CD3−CD19+CD20+CD27− naïve B cells or CD3−CD19+CD20+CD27+ antigen experienced B cells (mature) isolated from peripheral mononuclear cells. RosettaAntibody structural models were created by identifying homologous templates for the CDRs, VH–VL orientation, and FRs; assembling the templates into one model; *de novo* modeling the CDR-H3 loop; rigid-body docking the VH–VL interface; side-chain packing; and minimizing in the Rosetta energy function[37].

Since *de novo* modeling of long loops is challenging, DeKosky and Lungu *et al.* limited their antibody set to the more tractable subset of antibodies with CDR-H3 loop lengths under 16 residues. They compared their models for seven human germline antibodies with solved crystal structures and found models had under 1.4 Å backbone RMSD for the FR and under 2.4 Å backbone RMSD for the CDR-H3 loop.

I used the FIRST-PG method[39,43] to estimate flexibility from the RosettaAntibody homology models, determining the number of backbone DOFs for the CDR-H3 loop as each hydrogen bond is broken in order from weakest to strongest. FIRST models the antibody as a molecular graph where nodes represent atoms and edges represent atomic interactions. An extension of the PG algorithm uses this molecular graph to compute the DOFs of the CDR-H3 loop. To mitigate the effects of homology modeling inaccuracies on the FIRST-PG analysis, I used an ensemble of ten lowest-energy RosettaAntibody models. FIRST-PG analysis on structural ensembles has been shown to predict hydrogen–deuterium exchange and protein flexibility[40]. To account for varying CDR-H3 loop lengths, I scaled the calculated DOFs by a theoretical maximum value (Methods). Figure 4.1A shows a curve of the scaled DOFs averaged over all naïve or mature antibodies as a function of the hydrogen-bonding energy cutoff used in the FIRST-PG analysis. At a cutoff of 0 kcal/mol, all hydrogen bonds are intact and the average CDR-H3 loop scaled DOFs are about 20% of the theoretical maximum. Moving from right to left on the plot increases the minimum energy cutoff for including interactions in the FIRST graph; effectively hydrogen bonds of increasing strength are "broken" and the available DOFs rise from 20% to over 90% of the maximum theoretical flexibility while the loop becomes unstructured (unfolded) in FIRST.

I compared the DOFs distributions for naïve and mature antibodies at every hydrogen-bonding energy cutoff by two-sample Kolmogorov–Smirnov (KS) testing, with null hypothesis being that the two distributions are identical (Figure 4.1A). There is no difference in the average, scaled DOFs. To further quantify this comparison, I computed the average AUC plus-or-minus one standard deviation for both antibody sets. The average AUC values

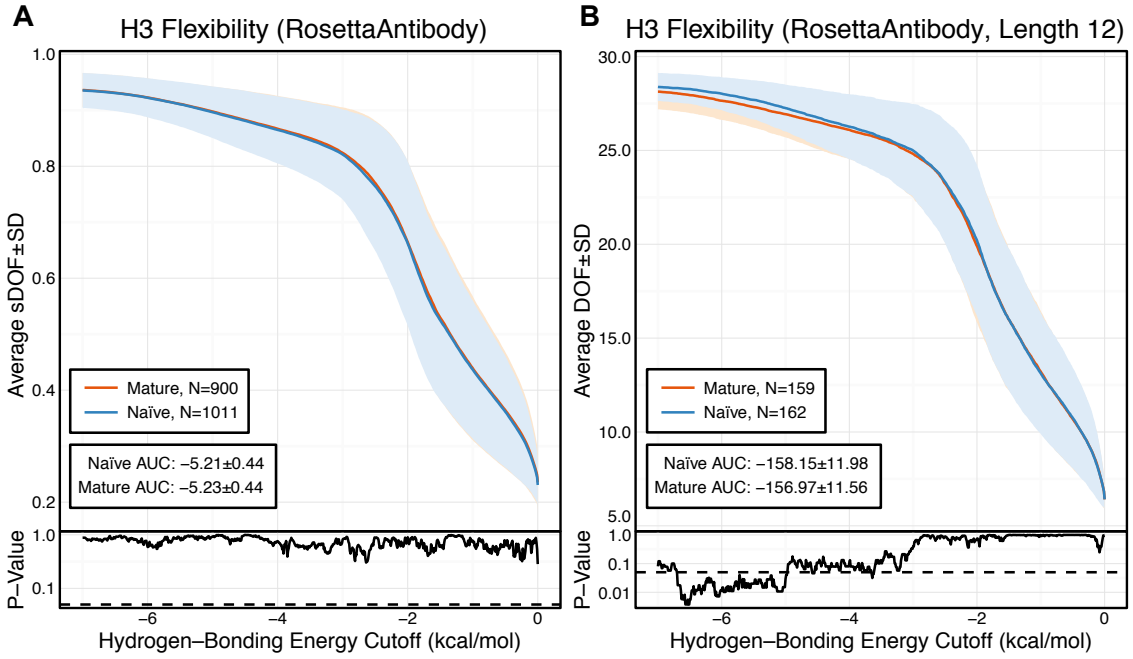

Figure 4.1: CDR-H3 loop flexibility analysis of the immunomic antibody set reveals that no difference in naïve (blue) and mature (red) antibodies. FIRST-PG was used to determine the degrees of freedoms (DOFs) as a function of hydrogen-bonding energy cutoff in RosettaAntibody models of the 1,911 most frequent public antibodies. Results were split, depending on whether the antibody was naïve or mature, as determined by B-cell surface receptors, and the mean DOFs were calculated along with the SD, shown in a lighter shade of the respective color. Subplots, below each main plot, show the p-value computed by a two-sample Kolmogorov–Smirnov (KS) test comparison of the naïve and mature DOFs distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same. A dashed line indicates a p-value of 0.05. (A) To permit comparison across loops of multiple lengths, the DOFs were scaled to a theoretical maximum for each length (a value of 1 indicates all DOFs are available, whereas a value of 0 indicates no DOFs are available). I found the scaled DOFs to be similar for both naïve and mature antibodies, quantified by the KS test p-values and area under the curve (AUC) $\pm$ SD: -5.21 ± 0.44 and -5.23 ± 0.44, respectively. (B) To exclude length effects on flexibility calculations, I compared DOFs for the most popular length (12 residues). I found the naïve AUC $\pm$ SD at 158.15 ± 11.98 and mature AUC $\pm$ SD at -156.97 ± 11.56 to be similar. The distributions appear similar at cutoffs between 0 and -5.0 kcal/mol, according to the KS test p-values.

are identical for the naïve (-5.21 ± 0.44) and mature antibody repertoires (-5.23 ± 0.44). This lack of difference persists (AUC -158.15 ± 11.98 [naïe] vs. -156.97 ± 11.56 [mature]) when accounting for CDR-H3 loop length, by comparing loops of only length 12, the most popular length (Figure 4.1B), and so the observed similarity of DOFs in naïve and mature antibodies is not due to averaging over loops of different lengths. Thus, on the immunomic repertoire scale, I do not observe the difference in flexibility between naïve and mature antibodies predicted by the paratope rigidification hypothesis.

Before amending the rigidification hypothesis in light of these results, I considered several alternative explanations for my observations. First, I addressed whether the use of homology models for flexibility analysis introduced inaccuracies by analyzing a large set of antibody crystal structures and Rosetta-generated models from that set with varying quality, ranging from models with sub-angstrom backbone RSMD to models that may be several angstroms off (and more representative of an average homology model). Next, I addressed whether backbone DOFs, as calculated by FIRST-PG, were a good measure of flexibility, by assessing flexibility through two alternative measures: B-factors and MD simulations. Additionally, I addressed whether averaging flexibilities and comparing across many germlines affected results, by detailed flexibility analysis of previously studied naïve–mature antibody pairs and RosettaAntibody-modeled pairs.

4.4.2 Only small flexibility differences are observed between naïve and mature antibodies in the crystallographic set

4.4.2.1 Preparation of an antibody crystal structure dataset

Of course, the strongest critique of the immunomic antibody set is that these models are only approximating the actual antibody structure. Thus, I applied FIRST-PG analysis to a large set of antibody crystal structures. I curated the set of all non-redundant mouse and human antibody crystal structures from SAbDab[45]. To be consistent with the models produced by RosettaAntibody, I truncated the structure of each antibody to only the Fv domain, excluding other antibody regions or antigen. Then, I used IMGT/3Dstructure-DB[46]

to identify the variable domain genes and determined the number of somatic mutations by aligning the sequence derived from the crystal structure to the IMGT-determined V-gene. I defined mature antibodies as those possessing at least one somatic mutation in either V-gene. The full dataset has 922 antibodies of which 23 are naïve.

4.4.2.2 FIRST-PG analysis of crystal structures

From the crystal structures, I created two sets of structural ensembles and assessed flexibility by FIRST-PG. Flexibility analysis has previously been shown to be more accurate on ensembles in comparison to analysis using single (snapshot) conformers[40,57]. Ensembles of ten representative structures were generated from the initial crystal structure by using either Rosetta FastRelax[48] or the refinement step of RosettaAntibody[37,38], which I term KIC ensembles after the loop modeling algorithm used in refinement[49]. Rosetta FastRelax samples structures around the crystallographic, local energy-minimum, with typically < 1 Å backbone RMSD, whereas the refinement step of RosettaAntibody samples a more diverse set of low-energy CDR-H3 loop conformations and VH–VL orientations. Thus, FastRelax ensembles are representative of the crystal structures, whereas KIC ensembles are representative of RosettAntibody homology models. By comparative FIRST-PG analysis of the two sets, I can assess the effects of modeling inaccuracies on flexibility analysis.

The scaled DOFs as calculated by FIRST-PG for FastRelax ensembles of antibody crystal structures are shown in Figure 4.2A. There are only minor differences between the naïve and mature flexibility curves, two-sample KS testing reveals insignificant p-values ($\gg 0.05$) for all hydrogen-bonding energy cutoffs, and the AUC is similar for both sets (4.70 ± 0.46 [naïve] vs. 4.70 ± 0.48 [mature]). Again, I considered the possibility that different distributions of loop lengths in the two sets obscures the affinity maturation contributions to flexibility. Therefore, I analyzed loops of length 10 (Figure 4.2B), the single most common length in the crystallographic set. When loops of a single length were compared, there was a separation between the naïve and mature sets, with the naïve antibody set average DOFs being

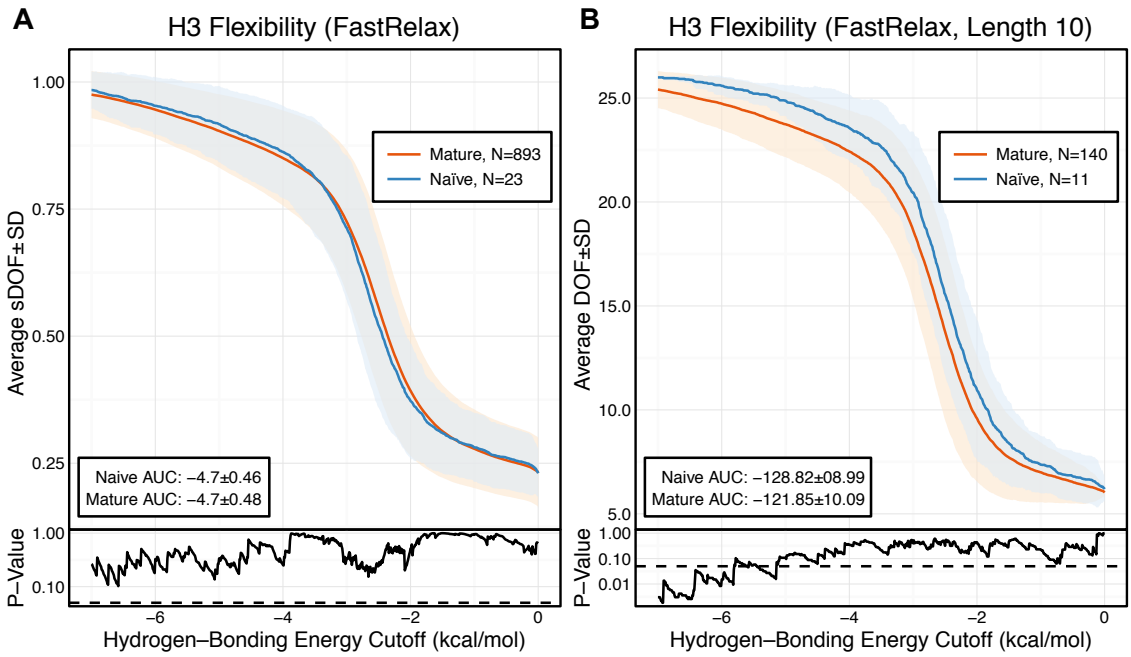

Figure 4.2: When accounting for length, CDR-H3 loop flexibility analysis of the crystallographic antibody set reveals naïve (blue) antibodies to be slightly more flexible than mature (red). FIRST-PG was used to determine the DOFs as a function of hydrogen-bonding energy cutoffs in crystal structure ensembles created by Rosetta FastRelax. Results were split, depending on whether the antibody was naïve or mature, as determined by BLAST alignment to its germline V-genes, and the mean DOFs were calculated along with the standard deviation, shown in a lighter shade of the respective color. Subplots, below each main plot, show the p-value computed by a KS-test comparison of the naïve and mature DOF distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same. A dashed line indicates a p-value of 0.05. (A) To permit comparison across loops of multiple lengths, the DOFs were scaled to a theoretical maximum for each length (a value of one indicates all DOFs are available whereas a value of zero indicates not DOFs are available). I found the scaled DOFs to be similar for both naïve and mature antibodies, quantified by KS-test p-values and the AUCs $\pm$ SD: -4.70 ± 0.46 and -4.70 ± 0.48, respectively. (B) To exclude length effects on flexibility calculations, I compared DOFs for the most popular length (10 residues). I found the naïve AUC $\pm$ SD at -128.82 ± 8.99 was greater than the mature AUC $\pm$ SD at -121.85 ± 10.09, but still within a standard deviation. The distributions appear similar at cutoffs between 0 and -6.0 kcal/mol, according to the KS-test p-values.

consistently greater than the mature set, but not significantly so, except for some energy cutoffs below -5 kcal/mol, according to KS testing. As expected, the AUC values differ, but are within a standard deviation (128.2 ± 9.0 [naïve] vs. 121.9 ± 10.1 [mature]). I repeated FIRST-PG analysis for KIC ensembles of antibody crystal structures and observed similar results (Supplementary Figure 4.A.1): for scaled DOFs, the AUC was 5.91 ± 0.20 (naïve) vs. -5.81 ± 0.26 (mature) and, for loops of length 10 only, the AUC was -154.10 ± 4.80 (naïve) vs. -150.44 ± 7.73 (mature). Thus, there does not appear to be a large, consistent CDR-H3 loop flexibility difference across all antibody crystal structures analyzed.

4.4.2.3 B-factor analysis of crystal structures

However, I have not accounted for the possibility that backbone DOFs as calculated by FIRST-PG may not capture the effects of affinity maturation on CDR-H3 loop flexibility. Thus, I assessed loop flexibility as determined by atomic temperature factors or B-factors. In protein crystal structures, B-factors measure the heterogeneity of atoms in the crystal lattice. Thus, rigid regions have lower B-factors as they are more homogeneous throughout the crystal whereas flexible regions have higher B-factors as they are less homogeneous throughout the crystal. B-factors are also affected by crystal resolution, so I cannot compare raw values across structures of varying resolution. Instead, I computed a normalized B-factor z-score, which has zero mean and unit standard deviation for each antibody chain. Finally, to account for different CDR-H3 loop lengths, I averaged the B-factor z-scores for the CDR-H3 loop residues.

Figure 4.3A shows the distributions of B-factor z-scores averaged over the CDR-H3 loop residues of naïve and mature antibodies. Both distributions span a similar range and overlap significantly, with the naïve curve peak shifted toward higher values than the mature. The majority of the naïve CDR-H3 loop B-factor z-score averages were positive (65%), whereas the majority of the mature CDR-H3 loop B-factor z-score averages were negative (64%). To address the question whether these distributions arose from the same

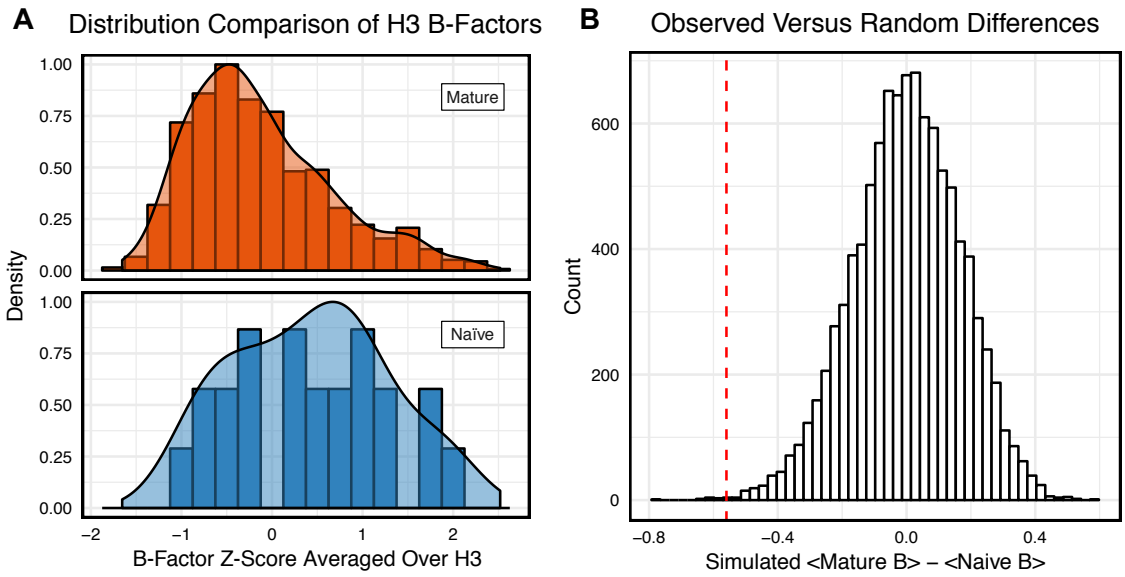

Figure 4.3: Comparison of the distribution of average CDR-H3 loop B-factor Z-scores in antibody crystal structures suggests that naïve are more flexible than mature. (A) Distributions of average CDR-H3 loop B-factors for the crystallographic set of antibodies are distinct for the mature (orange) and naïve (blue) sets. The mature antibody CDR-H3 loops have lower B-factors than the naïve, corresponding to more rigidity. Bars show binned counts in intervals of 0.25. Both the bars and smoothed densities are normalized so the maximum value is 1. A two-sample KS test confirms different underlying distributions with a p-value of 0.006 and maximum vertical deviation, D, of 0.36. (B) The observed difference in distribution means is difficult to replicate by random chance, occurring only 6.6 ± 2.6 times out of 10,000 simulations. Compare the observed difference in means (red line, dashed) to simulated differences (white bars) acquired by randomly assigning B-factor values from the original distributions to either a naïve or mature set, in the observed numbers (N_{mature} = 897 and N_{naive} = 23), before computing the difference in means.

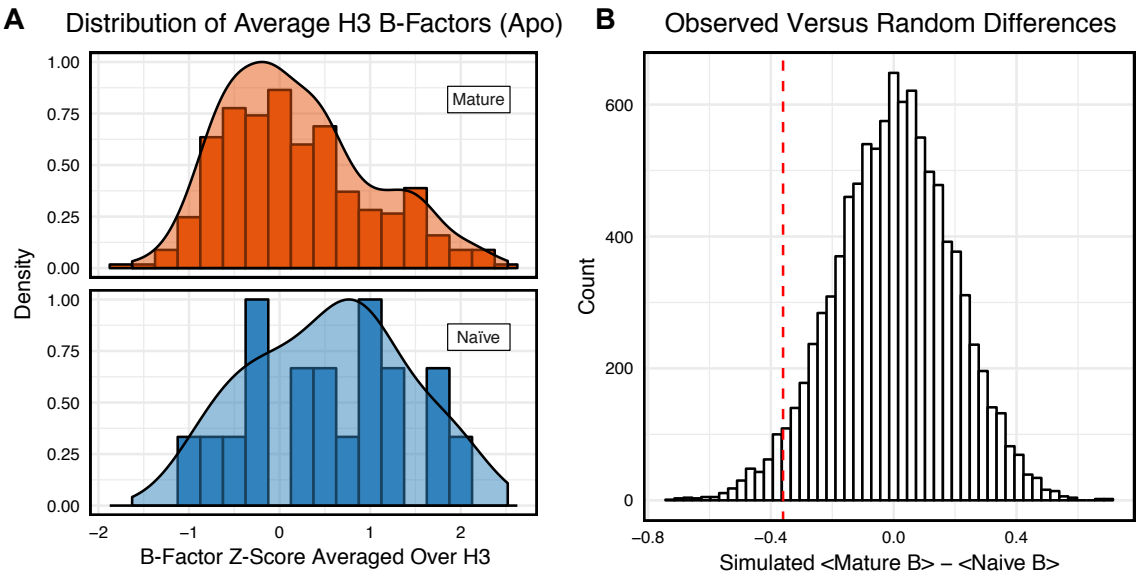

Figure 4.4: When considering only antigen-free crystal structures (to control for rigidification upon antigen biding), the difference between naïve and mature average CDR-H3 loop B-factor z-score distributions is small. (A) The distributions of CDR-H3 loop average B-factors are less distinct between the mature (orange) and naïve (blue) sets. Bars show binned counts in intervals of 0.25. Both the bars and smoothed densities are normalized so the maximum value is 1. A two-sample KS test results in a p-value of 0.15 and D of 0.27, indicating that the null hypothesis of indistinguishable underlying distributions cannot be discarded. (B) The observed difference in distribution means (red line, dashed) is occasionally replicated in random resampling (white bars). When average CDR-H3 loop B-factor z-scores are pooled and randomly assigned to either a naïve or mature set, in the observed numbers ($N_{mature} = 355$ and $N_{naive} = 18$), the observed difference in means is matched or surpassed in 340 ± 20 out of 10,000 simulated differences.

underlying distribution I turned to randomization testing, as described in the Methods. The observed difference in distribution means is matched by only $0.066 \pm 0.026\%$ of simulated differences (Figure 4.3B), indicating that naïve and mature distributions are likely distinct. Furthermore, a two-sample KS test confirms the distributions to be distinct, with a maximum vertical deviation, D, of 0.36 and a p-value of 0.006.

However, I was concerned that the mixing of bound and unbound crystal structures would influence results, as I previously observed bound structures to have lower average B-factors[58]. Furthermore, in the PDB-derived dataset, naïve antibodies were mostly crystallized in the unbound state (19 of 23), whereas mature antibodies were mostly co-crystallized with their cognate antigen (544 of 899). In conjunction, these two observations suggested that the high number of antigen-bound mature antibody crystal structures was the primary

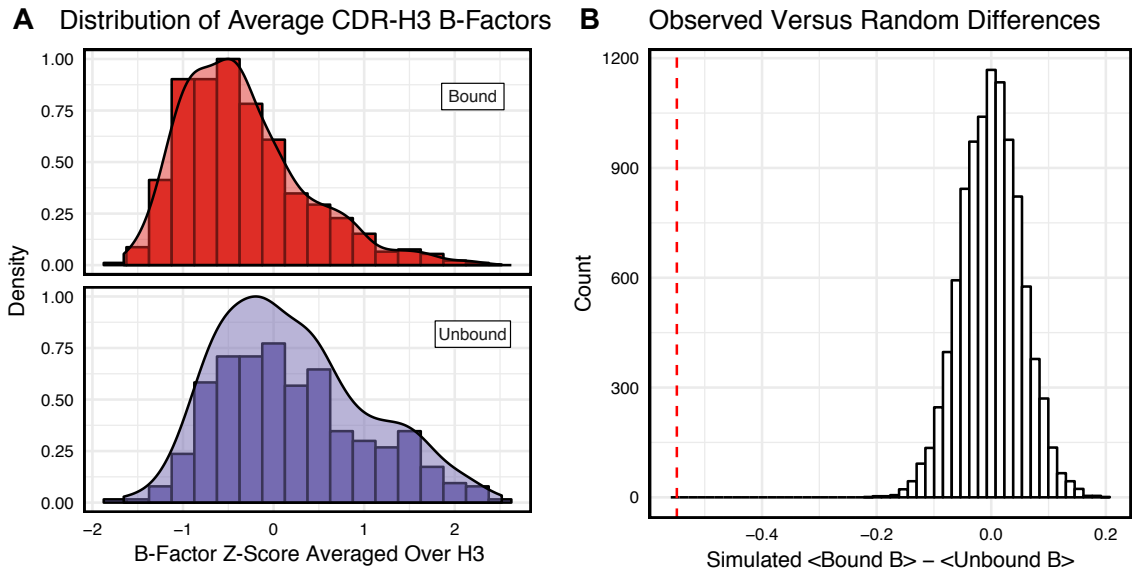

Figure 4.5: Antigen-bound and antigen-free distributions of B-factor z-scores are distinct. (A) Distributions of CDR-H3 loop average B-factors for the crystallographic set of antibodies are distinct for the antigen-bound (red) and antigen-free (purple) sets. Bound antibody CDR-H3 loops have lower B-factors than unbound, corresponding to more rigidity. Bars show binned counts in intervals of 0.25. Both the bars and smoothed densities are normalized so the maximum value is 1. Distributions appear distinct according to a two-sample KS test with a p-value of 2.2E-16 and D of 0.31. (B) The observed difference in distribution means (red line, dashed) is never replicated in 10,000 attempts at random resampling (white bars). Simulated differences were acquired by randomly assigning values from both sets to either a naïve or mature set, in the observed numbers ($N_{bound} = 546$ and $N_{naive} = 374$), before computing the difference means.

driver of the difference between naïve and mature B-factor z-scores. Thus, I compared the B-factor averages of unbound structures only and found that while the distributions appear to be distinct (Figure 4.4A), when the difference in distribution means is compared to a randomized set, $3.4 \pm 0.2\%$ of random differences are greater than or equal to the observed differences, and the distributions fail a two-sample KS test ($D = 0.27$, $p = 0.15$). Thus, the difference between naïve and mature antigen-free crystal structures does not appear significant.

As I conjectured, a significant difference was found between the bound and unbound distributions (Figure 4.5), with a two-sample KS test confirming the difference between the distributions ($D = 0.31$, $p < 2.16E-16$) and randomized testing never showing a difference in means as large as the observed difference. Additionally, I considered other possible origins of difference between the naïve and mature distributions that are not related

to affinity maturation, including comparison across species, crystal structure resolutions, CDR-H3 loop lengths, and if the CDR-H3 loop was at a crystal contact or not. I found none of these to have as clear of an effect on the distribution of B-factor averages as whether or not antigen was bound (Supplementary Figures 4.A.2 and 4.A.3). In summary, the distributions of B-factor z-score averages (Figures 4.3, 4.4, 4.5) suggest that both the naïve and mature antibody sets possess CDR-H3 loops of varying flexibility and that neither set is significantly more flexible or rigid than the other.

4.4.3 Comparison of mature to naïve-reverted models reveals varying rigidification across matched pairs

Having not observed consistent rigidification of the CDR-H3 loop in two large sets of antibodies, I postulated that rigidification was not a repertoire-wide phenomenon (i.e. all mature antibodies are not more rigid than all naïve antibodies), but it could still be plausible that matched pairs of naïve and mature antibodies would reveal rigidification.

To investigate this hypothesis, I selected ten mature antibodies from the SAbDab set with CDR-H3 loops of length 10, a length for which loop modeling performs well[49,59]. I identified antibodies that had at least 5 (97% sequence identity), but no more than 25 (85% sequence identity), mutations when compared to the germline V-genes. To control for species, half of the selected antibodies were human and half were mouse. I reverted the mature antibody sequences to naïve using the germline sequences from the aligned V-genes, as described in the methods, and using germline J-genes from sequence alignments from IMGT/DomainGapAlign[46]. The reverted sequences are reported in the Supplemental Material. I then used RosettaAntibody to generate homology models for the naïve-reverted sequences. I analyzed the ensembles of the ten lowest-energy homology models using FIRST-PG. To ensure fair comparison, I also used FIRST-PG to analyze homology model ensembles of the mature sequences. To provide an estimate for the accuracy of RosettaAntibody homology models, I computed RMSDs for the mature models using the known crystal structures and found all had sub-2-Å CDR-H3 loop backbone RMSD, calculated after

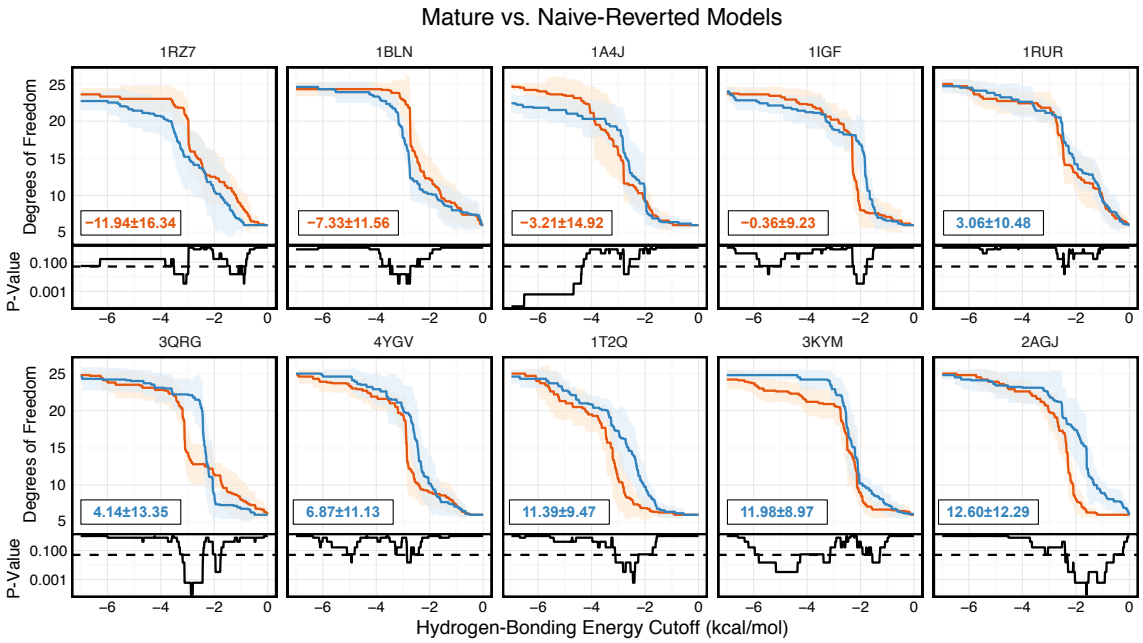

Figure 4.6: FIRST-PG analysis of ten RosettaAntibody-modeled mature/naïve-reverted antibody pairs (CDR-H3 loop length of 10 residues) shows that affinity maturation does not always result in CDR-H3 loop rigidification. Naïve values are colored blue, while mature values are color red. The difference between mature and naïve AUCs is reported in the bottom left of each sub-figure, with a positive value indicate a more flexible naïve antibody. Four out of the ten cases have mature antibodies with AUC greater than their naïve counterparts. Subplots, below each main plot, show the p-value computed by a KS-test comparison of the naïve and mature DOF distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same and a dashed line indicating a p-value of 0.05.

alignment of the heavy chain FR, with 7 of 10 antibodies having sub-Å RMSD.

Of the ten naïve/mature antibody pairs I analyzed, six showed a decrease in flexibility and four showed an increase in flexibility upon affinity maturation (Figure 4.6). These ten antibodies demonstrate the breadth of possible affinity maturation effects, from an expected flexibility decrease in antibody 2AGJ, with AUC decreasing by 9.34%, to the unexpected flexibility increase in antibody 1RZ7, with AUC increasing by 10.65%.

4.4.4 Analysis of 48G7 antibody

Having analyzed 1911 models, 922 crystal structures, and 10 paired-reverted models, I had yet to observe a consistent difference in CDR-H3 loop flexibility between naïve and mature antibodies, as previously reported in literature. Thus, I turned to three previously-studied

antibodies with known crystal structures and measured CDR-H3 loop flexibility. These are (1) the esterolytic antibody 48G7[15,31,32,34], (2) the anti-fluorescein antibody 4-4-20[22,25–27,30,32], and (3) a broadly neutralizing influenza virus antibody[21]. For all three antibodies, the effects of affinity maturation on CDR-H3 loop flexibility have been previously studied by both experiment and simulation, allowing comparison with my results. For brevity, I presently discuss the 48G7 antibody here, and full results for all antibodies are available in the Supplementary Material.

The 48G7 antibody was first studied through crystallography, with structures capturing the bound (holo) and unbound (apo) states of both the naïve and mature antibody[15]. Comparison between the naïve and mature CDR loop motions from the free to the bound state revealed minor changes, with the mature CDR-H3 loop being slightly more rigid and moving an Ångström less than the naïve upon antigen binding (Supplementary Figures 4.A.4). For each of the four crystal structures, I extracted B-factors and computed B-factor z-scores for the CDR-H3 loop, measuring the distance from the B-factor mean in standard deviations. B-factor z-scores for the CDR-H3 loop of apo-48G7 are shown in Figure 4.7A. The mature antibody has lower B-factors than the naïve antibody throughout the entire CDR-H3 loop. This observation also holds for the holo-48G7 antibody structures as well (Supplementary Figure 4.A.5). Supplementary Table 4.A.1 summarizes B-factors averaged over the whole CDR-H3 loop. These B-factor results agree with the prior crystallographic observations.

Follow-up studies on 48G7 used MD simulations to assess flexibility. Briefly, 500 ps short MD simulations of the naïve and mature antibodies in the presence of antigen with an explicit solvent model (TIP3P) found the CDR-H3 loop to be more flexible in the naïve than in the mature antibody by comparison of RMSFs[29], but 15 ns MD simulations of the naïve and mature antibodies in the absence of antigen with an implicit solvent model (GB/SA) found no difference between the two, again by comparison of RMSFs[31]. Another study based on an elastic network model also suggested that, in the absence of

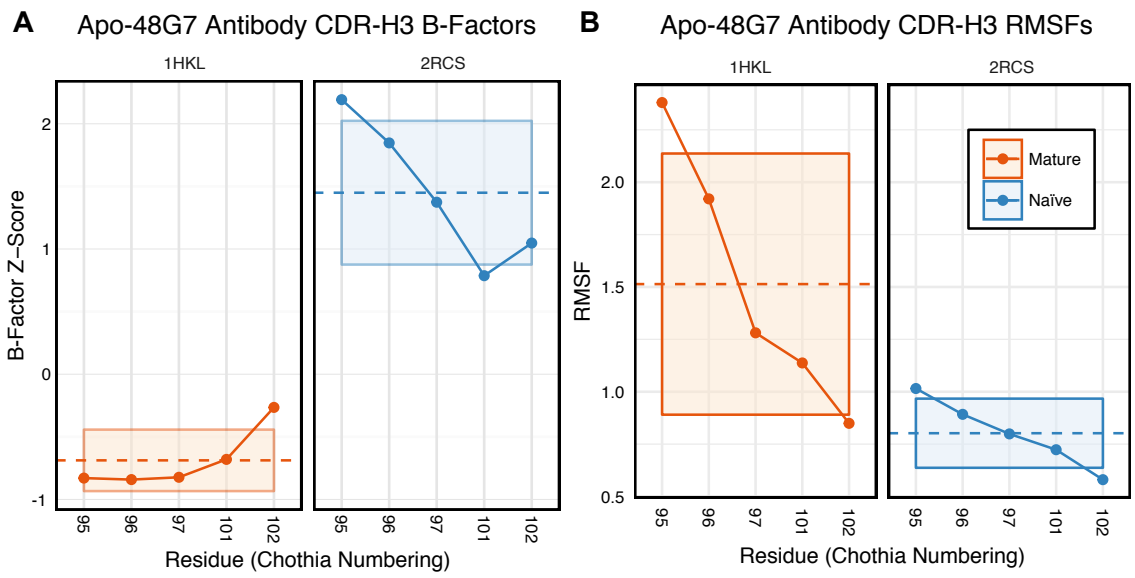

Figure 4.7: Analysis of catalytic antibody 48G7 by CDR-H3 loop B-factors and RMSFs shows conflicting results. (A) Comparison of normalized B-factor values for the CDR-H3 loop of the 48G7 antibody in crystal structures of the unbound naïve (dark blue) and mature (dark orange) antibodies reveals a more rigidity in the mature antibody. The dashed line indicates the average value and is outlined by a box defined by the average plus-or-minus the standard deviation. (B) Comparison of CDR-H3 loop RMSFs for the MD simulations of the naïve and mature 48G7 antibodies shows the opposite.

antigen, the fluctuations of the naïve and mature 48G7 were similar, but their binding mechanisms could differ depending on response to antigen binding; the naïve antibody shows a discrete conformational change induced by antigen whereas the mature antibody shows lock-and-key binding[60]. Due to the contentious nature of these results, 200 ns MD simulations were run for the 48G7 naïve and mature antibodies in the absence of antigen with an explicit solvent model (TIP3P). I measured both RMSDs and RMSFs for the CÎś atoms along the CDR-H3 loop and computed the difference between the naïve and mature antibodies (Supplementary Table 4.A.1). Figure 4.7B shows that the CDR-H3 loop RMSFs are consistently greater for the mature than the naïve 48G7 antibody.

Finally, as I have done through this study, I used FIRST-PG to measure CDR-H3 loop flexibility. To limit the effects of crystal structure artifacts on FIRST-PG analysis, I used an ensemble of ten representative structures, derived by clustering trajectory frames and selecting ten structurally distinct cluster medians from the MD simulations, similar to a

previous flexibility study for this antibody[32]. The CDR-H3 loop flexibility of apo-48G7, as determined by FIRST-PG analysis of MD ensembles is shown in Figure 4.8. The FIRST-PG analysis showed no significant difference between the mature and naïve antibodies.

In addition to using MD simulations to generate ensembles, I used ensembles generated by RosettaAntibody and Rosetta FastRelax, permitting direct comparison. The CDR-H3 loop flexibility of apo-48G7, determined by FIRST-PG analysis of FastRelax and RosettaAntibody ensembles, is shown in Figure 8. The curves from FastRelax and the MD simulation are similar for low-energy cutoffs (e.g. in the range of 0.0 to -3.0 kcal/mol), with the naïve and mature DOFs being the same. These curves diverge at higher-energy cutoffs where the FastRelax curve shows a more flexible naïve antibody and the MD curve does not. The curve from RosettaAntibody ensembles differs from the two and shows a more flexible mature antibody at low-energy cutoffs and a more flexible naïve at high-energy cutoffs. For less visual and more quantitative comparisons, I computed the AUC of the DOF versus hydrogen-bonding energy cutoff plots (Supplementary Table 4.A.1). I find the AUC is only slightly greater for naïve than mature antibodies in the FastRelax and RosettaAntibody ensembles, with the naïve AUC reducing by only 3.9% and 0.2%, respectively, upon maturation. MD ensembles show the opposite outcome, with the mature antibody having 1.3% greater AUC than the naïve.

Further validation was carried out on two other previously studied antibodies and reported in the Supplementary Table 4.A.1 and Supplementary Figures 4.A.5 and 4.A.6. For the 4-4-20 antibody, antigen-bound structures were compared and the average mature B-factors were within a standard deviation of the naïve. For the influenza antibody, average B-factors were compared between an unbound naïve and a bound mature crystal structure, showing significant rigidification. However, results are conflated due to the lack of unbound crystal structures, as in bound structures antibody–antigen contacts artificially increase rigidity of the CDR-H3 loop. In contrast to B-factor analyses, FIRST-PG analyses yielded mixed results for these two antibodies. The 4-4-20 antibody was found to become more

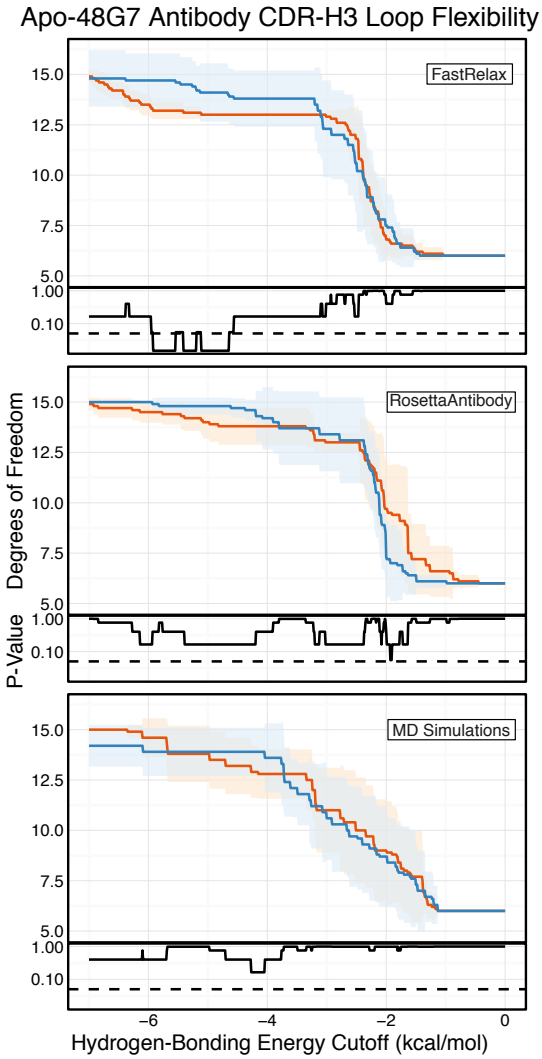

Figure 4.8: FIRST-PG analysis of naïve (dark blue) and mature (dark orange) 48G7 antibodies using either Rosetta FastRelax-, RosettaAntibody-, or MD-generated 10-member ensembles does not show a difference between the naïve and mature antibodies. FIRST-PG analysis calculates the DOFs of CDR-H3 loop as a function of hydrogen-bonding energy cutoff. Subplots, below each main plot, show the p-value computed by a KS-test comparison of the naïve and mature DOF distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same and a dashed line indicating a p-value of 0.05. While the FastRelax ensembles appear distinct in the range of -6 to -3 kcal/mol, the naïve and mature are indistinguishable for both the RosettaAntibody and MD ensembles.

flexible upon maturation by FIRST-PG analysis of all but Rosetta KIC ensembles. The influenza antibody was found to become more rigid upon mature by FIRST-PG analysis of all but Rosetta FastRelax ensembles. Finally, I analyzed RMSDs and RMSFs from MD simulations and found that the mature 4-4-20 antibody has higher CDR-H3 loop RMSD, but lower RMSF, values than the naïve while the mature influenza antibody was found to have lower values for both (Supplementary Table 4.A.1). As with the repertoire analysis, I do not see consistent rigidification in previously studied antibodies. I consider the significance of this result and compare my analysis in detail to past analyses of flexibility in the Discussion section.

4.5 Discussion

4.5.1 The varying effects of affinity maturation on CDR-H3 flexibility

Affinity maturation, through a series of somatic hypermutation events and selection processes, can evolve a low-affinity, naïve antibody to bind an antigen with both high affinity and specificity. Elucidating the affinity maturation process is desirable to understand molecular evolution, develop antibody engineering methods, and guide vaccine development[61]. Past studies have suggested that, with few exceptions[28], naïve antibodies are highly flexible and maturation leads to improved affinity and specificity through the optimization and rigidification of the antibody paratope, and in particular the CDR-H3 loop[21,26,30–32]. However, these studies have been limited, often focusing on a single antibody and assessing flexibility indirectly. I sought to test the generalizability of the rigidification-upon-maturation hypothesis. I was enabled by the large number of antibody structures in the PDB, homology models generated from high-throughput repertoire sequencing data, and the FIRST-PG method for rapid structural flexibility calculation to ask whether affinity maturation leads to CDR-H3 loop rigidification.

Unexpectedly, in a comparison of flexibility of repertoires, the data show little difference between naïve and mature antibodies: FIRST-PG calculations showed no difference for

RosettaAntibody homology model ensembles of the most common naïve and mature antibodies in human peripheral blood cells. The same calculations showed no difference in CDR-H3 loop DOFs of crystal structures under two different refinement schemes (FastRelax and KIC). After accounting for the presence/absence of antigen, CDR-H3 loop B-factor distributions were similar for both mature and naïve antibody crystal structures. These results indicate that rigidification of the CDR-H3 loop does not always occur upon affinity maturation.

Since these observations did not indicate clear rigidification over two sets of antibodies, I considered the following possibilities: (1) comparison of different length CDR-H3 loops was unfair because longer loops are inherently more flexible, (2) comparison of different antibodies was unfair because different combinations of gene segments and VH–VL pairs will result in different flexibilities, (3) mutations within CDR-H3 loop, which I could not identify for the PDB set because of the difficulty in D/J-gene alignments, may have modulated flexibilities of CDR-H3, (4) inaccuracies in the computational methods could preclude observation of rigidification, and (5) FIRST-PG-measured backbone DOFs are not a good measure of flexibility. To address the first concern, I analyzed loops of consistent length via B-factor and FIRST-PG (Figures 4.1B & 4.2B, Supplementary Figures 4.A.1 & 4.A.2). I found that, according to KS testing and when accounting for the presence/absence of antigen, B-factor distributions were not distinct for naïve and mature sets of antibodies with same length CDR-H3 loops (length 10 for the crystallographic set and 12 for the repertoire model set). I also found that FIRST-PG DOFs AUC values of the naïve and mature sets of antibodies with the same length CDR-H3 loops were within a standard deviation for RosettaAntibody, FastRelax, and KIC ensembles. So, even when accounting for length, mature antibodies are not significantly more rigid than naïve ones.

To address the concern that comparison of sets of antibodies originating from different VH and VL genes is unfair, I analyzed mature/naïve antibody pairs that had been previously studied and mature/naïve-reverted pairs that I generated with RosettaAntibody

and analyzed by FIRST-PG (Figures 4.6, 4.7, 4.8, Supplementary Table 4.A.1). I found that CDR-H3 loop B-factors did not always indicate rigidification upon maturation and for the 7G12 antibody I observed the reverse effect (Supplementary Figure 4.A.7). I also found that mature antibodies did not always become more flexible upon naïve reversion, but instead displayed a breadth of behaviors (Figure 4.6). So, when analyzing matched naïve/mature pairs, I do not see consistent rigidification upon maturation.

My analysis of previously studied naïve/mature antibody pairs coupled with the earlier repertoire analysis should alleviate concerns that the flexibility results for the PDB set were strongly affected by the inability to align D/J-gene segments and thus consider mutations in the CDR-H3 loop. The previously studied pairs included CDR-H3 mutations and the repertoire set had antibody sequences determined by Illumina MiSeq sequencing with naïve/mature status assigned by the absence/presence of the CD27 cell-surface receptor. In both cases, the naïve and mature sequences were determined through the entire Fv, and flexibility analysis still revealed mixed results.

Finally, to address the concern that RosettaAntibody models may not be accurate enough to be useful for FIRST-PG calculations, I tested FIRST-PG on a range of structural ensembles with varying deviation from the crystal structure. I found no difference in the naïve vs. mature antibody CDR-H3 loop AUC of the FIRST-PG results, regardless of the ensemble generation method used (compare Figure 4.2 and Supplementary Figure 4.A.1). I also determined flexibility through alternative measures such as crystal structure B-factors and RMSFs in MD simulations. For both, affinity maturation was not found to have a consistent, rigidifying effect. Thus, even if model inaccuracies confound analysis, other data support the same hypothesis.

4.5.2 Comparison with prior results

My analysis included several antibodies that have been the subject of previous flexibility studies, permitting a direct comparison. One of the most studied antibodies is the

anti-fluorescein antibody, 4-4-20. Spectroscopic experiments measuring the response of a fluorescent probe (fluorescein) and MD simulations measuring CÎś atom fluctuations suggested that somatic mutations restrict conformational fluctuations in the mature antibody[25,27,30]. My analysis of 4-4-20 was not as clear: I observed no significant difference in naïve vs. mature CDR-H3 loop crystallographic B-factors (Supplementary Figure 4.A.5) and found the mature antibody to be more rigid in FIRST-PG calculations only in the $-2.0 - -0.0$ kcal/mol range of hydrogen-bonding energy cutoffs (Supplemental Figure 4.A.6). Similar mixed results were observed by Li *et al.*[32] who used a Distance Constraint Model (DCM) to analyze flexibility in an ensemble of 4-4-20 conformations drawn from MD simulations. They found increases in structural rigidity of the CDR-H3 loop, as determined by the DCM, occurred upon affinity maturation, but these increases did not correspond to decreases in dynamic conformational fluctuations, as determined by RMSFs from MD simulations. Further studies artificially matured 4-4-20 by directed evolution, resulting in a femtomolar-affinity antibody, 4M5.3[62], but the crystal structures of 4M5.3 and 4-4-20 were almost identical (the reported backbone RMSD is 0.60 Å) and thermodynamic measurements suggested that the affinity improvement was achieved primarily through the enthalpic interactions with subtle conformational changes[63]. This observation was contradicted by Fukunishi *et al.*[64], who performed steered MD simulations to analyze the effects of the mutations on the flexibility of 4-4-20 and 4M5.3. By applying external pulling forces between the antibodies and the antigen along a reaction coordinate, they quantified the interactions and showed that, during the simulations, fluctuations of the antibody, especially the CDR-H3 loop, and of the antigen were indeed larger in 4-4-20 than in the more matured antibody, 4M5.3[64]. Thus, there is some variation not only in these results, but also in the literature as to the effects of affinity maturation on 4-4-20.

Another set of well-studied antibodies are the four catalytic antibodies: 48G7, 7G12, 28B4, and AZ-28. In fact, the first crystallography studies to suggest rigidification of the CDR-H3 loop as a consequence of affinity maturation were performed on 48G7. Wedemayer

et al. observed larger structural rearrangements upon antigen binding in the CDR-H3 loop for the naïve antibody than the mature antibody (Supplementary Figure 4.A.4)[15]. Crystallization of the naïve unbound, naïve bound, mature unbound, and mature bound states for 7G12, 28B4, and AZ-28 revealed similar results[17,18]. Additionally, MD simulations of the four catalytic antibodies in implicit solvent were used to calculate CDR Cα atom B-factors[31]. Wong *et al.* showed a decrease in mature CDR-H3 loop B-factors in three cases (7G12, 28B4, and AZ-28) whereas no significant difference was observed for 48G7 (see Figure 2 in Wong *et al.*). Furthermore, for 48G7, Li *et al.* used MD simulation to generate structural ensembles and DCM analysis to determine flexibility. They found that the mature CDR-H3 loop is more rigid than the naïve, according to DCM, but used an unusual loop definition that included five additional flanking residues (see Fig. 1 in Li *et al.*), making comparison challenging (longer loops will be inherently more flexible), and they observed increases in the mature CDR-H3 loop RMSFs (see Fig. 8 in Li *et al.*)[32]. My analysis of CDR-H3 loop B-factors showed rigidification upon maturation for some of the 48G7 and 28B4 crystal structures (Figure 4.7 and Supplemental Figure 4.A.7), but not for 7G12 and AZ-28 structures (Supplemental Figures 4.A.7 & 4.A.8). FIRST-PG analysis of FastRelax, RosettaAntibody, and MD ensembles for 48G7 showed slight to no rigidification (Figure 4.8). Additionally, RMSFs from MD simulations for 48G7 showed higher values for the mature loop, contrary to the expectation that it is more rigid. My mixed results for the effects of affinity maturation on 48G7 are consistent with literature, but there is variation between my results and the literature as to the effects of affinity maturation on the other catalytic antibodies.

Finally, Schmidt *et al.* used X-ray crystallography, MD simulations, and thermodynamics measurements to investigate how somatic mutations affected the binding mechanism of anti-influenza antibodies[21]. They identified three mature antibodies, their unmutated common ancestor (UCA), and a common intermediate, all derived from a subject immunized with an influenza vaccine. The affinities of the mature antibodies were about 200-fold better

than the UCA. MD simulations of the UCA and the mature antibodies showed that CDR-H3 loop of the UCA could sample more diverse conformations than the mature antibodies, whose CDR-H3 loop sampled only conformations optimal for antigen binding, supporting the hypothesis that somatic mutations rigidify antibody structures. In another study by the same group[65], further MD simulations were performed on the same systems, showing that, although many somatic mutations typically accumulate in broadly neutralizing antibodies during maturation, only a handful of mutations substantially stabilize CDR-H3 loop and hence enhance the affinity of the antibodies for antigen. In my study, all the results (Supplemental Figures 4.A.5 and 4.A.6, Supplemental Table 4.A.1) for the anti-influenza antibody, except FIRST-PG flexibility calculations for the Rosetta FastRelax ensemble, show rigidification of the CDR-H3 loop as an effect of affinity maturation and agree with the detailed analysis of Schmidt *et al*.

For the three antibody families I analyzed in detail, I observed mixed effects of affinity maturation on two (catalytic antibodies and 4-4-20) and clear rigidification in one (anti-influenza antibody). For the two with mixed results, I note that past work has also shown conflicting results. I interpret these results as supportive of my repertoire-wide analysis that affinity maturation does not always rigidify the CDR-H3 loop.

4.5.3 Biophysical properties underlying antibody binding

Why is antibody CDR-H3 loop rigidification not a consistent result of affinity maturation? Consider the process of affinity maturation, which selects for antibody–antigen binding and against interactions with self or damaged antibodies (i.e. when deleterious mutations are introduced by activation-induced cytidine deaminase)[66]. Under these selection pressures, what is the benefit of CDR-H3 loop rigidification? Loop rigidification can only decrease the protein-entropy cost for antibody–antigen binding, having ostensibly no effect on enthalpy and solvent entropy of binding, and self-interactions. If CDR-H3 loop rigidification is just one of many biophysical mechanisms that can be selected for during affinity maturation,

then I do not expect to observe it consistently, in line with my results.

What are the other possible mechanisms then? Collectively, studies have shown that improved antibody affinity and specificity for antigen can be achieved by introducing additional interfacial interactions including hydrogen bonds, salt bridges, and van der Waals contacts[15,67–69]; increasing the buried surface area, either polar or apolar, depending on the antigen[19]; and improving interface shape complementarity[58], in addition to rigidification of the paratope[21]. A detailed review on the structural basis of antibody affinity maturation was recently published by Mishra and Mariuzza[70].

An interesting example of the consequences of the biological antibody selection process is the anti-hapten antibody, SPE7[71]. For SPE7, mutations leading to multi-specificity or promiscuity were beneficial–antibodies are multivalent, so an antibody capable of binding multiple antigens with intermediate affinity can gain an effective advantage through cooperative binding over an antibody capable of binding only one antigen. Crystal structures of SPE7 with different antigens and in its apo-state demonstrated that SPE7 can assume different conformations. Motivated by these observations, Wang *et al.* exploited MD simulations to investigate the binding mechanisms of SPE7[72]. The MD simulations and subsequent analyses suggested that multi-specific antigen binding is mediated by a combined mechanism of conformer selection and induced fit. Similar behavior, where the mature antibody is more flexible than the naïve has been observed for an antibody that recognizes the tumor-associated ganglioside GD2[73]. Such antibodies could not have arisen if CDR-H3 loop rigidification were a consistent result of affinity maturation.

4.6 Conclusions

I have conducted the largest-scale flexibility study of antibody CDR-H3 loops, analyzing 9,22 crystal structures and 1,911 homology models. I used B-factors and FIRST-PG to assess flexibility. I sought to identify the effects of affinity maturation on CDR-H3 loop flexibility, expecting the CDR-H3 loop to rigidify. I found that there were no differences in

the CDR-H3 loop B-factor distributions or FIRST-PG DOFs for naïve vs. mature antibody crystal structures and in the CDR-H3 FIRST-PG DOFs for homology models of repertoires of naïve and mature antibodies. These findings suggest that there is no general difference between naïve and mature antibody CDR-H3 loop flexibility in repertoires of naïve and mature antibodies. However, I observed rigidification of the CDR-H3 loop for some, but not all, antibodies when the mature antibodies were compared directly to their germline predecessors. Thus, I conclude that increased rigidity occurs alongside other affinity-increasing changes, such as improved interfacial interactions, increased buried surface area, and improved shape complementarity.

Further work must be done to address the issues observed here, i.e. inconsistent results across the different methods used to measure flexibility. One possible route is to explore experimental methods that directly measure protein dynamics across several timescales, and use them to study a relatively large (more than one or two antibodies) and diverse (e.g. from different source organisms or capable of binding different antigens) set of antibodies. For example, HDX-MS is capable of identifying protein regions with dynamics on timescales from milliseconds to days, has been previously used to study antibody dynamics, and has been correlated to FIRST-PG[28,40].

Finally, I note the need for more rapid and accurate antibody modeling methods. With the advent of high-throughput sequencing, there now exits a plethora of antibody sequence data, but little structural data. Accurate modeling can overcome the lack of high-throughput structure determination method and provide crucial structural data. These structures can then be used to address scientific questions on a larger scale than before, on the scale of the human antibody repertoire.

4.A Appendix

4.A.1 Rosetta modeling of crystals

Rosetta version 2017.26-dev59567 was used for all simulations. Antibody Fv regions were relaxed with the following command and options:

```
relax.linuxgccrelease -l pdb.list -ex1 -ex2 -use_input_sc -beta -
   nstruct 10
```

Antibody Fv regions had their CDR-H3 loop remodeled and relative VH–VL orientation resampled with the command and options below.

```
antibody_H3.linuxgccrelease -l pdb.list -ex1 -ex2 -nstruct 10 @abH3.
   flags
```

where, abH3.flags is a file containing the following additional options:

```
-antibody::remodel perturb_kic
-antibody::snugfit true
-antibody::refine  refine_kic
-antibody::cter_insert false
-antibody::flank_residue_min true
-antibody::bad_nter false
-antibody::h3_filter false
-antibody::h3_filter_tolerance 5

-extrachi_cutoff 0
-loops:legacy_kic false
-loops:kic_min_after_repack true
-loops:kic_omega_sampling
-loops:allow_omega_move true
-kic_bump_overlap_factor 0.36
-loops:ramp_fa_rep -loops:ramp_rama
-loops:refine_outer_cycles 2
-loops:max_inner_cycles 20
```

4.A.2 Rosetta modeling of sequences

Antibody Fv homology models were generate with RosettaAntibody in three steps: (1) assembly of the homologous components, (2) FastRelax of the grafted model, (3) CDR-H3 loop modeling and VH–VL docking. Homologous components were selected and assembled with the following command and options:

```
antibody.macosclangrelease -fasta pdb.fasta -antibody:
   n_multi_templates 1 -antibody:no_relax
```

The resulting "model-0.pdb" was the relaxed with constraints by:

```
relax.macosclangrelease -s model-0.pdb -flip_HNQ -no_optH false -relax
   :fast -relax:constrain_relax_to_start_coords -relax:
   ramp_constraints false -use_input_sc -ex1 -ex2 -nstruct 1
```

Finally, CDR-H3 loop modeling and docking of VH–VL was done by:

```
antibody_H3.linuxgccrelease -s grafting/model-0_0001.pdb -nstruct 1000
    @abH3.flags
```

with the following abH3.flags:

```
-antibody::remodel perturb_kic
-antibody::snugfit true
-antibody::refine refine_kic
-antibody::cter_insert false
-antibody::flank_residue_min true
-antibody::bad_nter false
-antibody::h3_filter false
-antibody::h3_filter_tolerance 5
-antibody:constrain_vlvh_qq

-ex1
-ex2
-extrachi_cutoff 0

-loops:legacy_kic false
-loops:kic_min_after_repack true
-loops:kic_omega_sampling 3
-loops:allow_omega_move true
-kic_bump_overlap_factor 0.36
-loops:ramp_fa_rep
-loops:ramp_rama
-loops:refine_outer_cycles 5
```

4.A.3 Reverted sequences

Mature sequences were aligned to germline V-genes as described in the methods. Additionally, sequences were aligned to germline J-genes using IMGT/DomainGapAlign, which yields germline alignments for both V- and J-genes. For example, the alignment of the variable region of 1T2Q can be extracted from: http://www.imgt.org/3Dstructure-

DB/cgi/details.cgi?pdbcode=1t2q. The germline sequence was used when possible. The alignments are shown below with the mature sequence above and naive below.

```
1A4J Heavy
QVQLLESGPELKKPGETVKISCKASGYTFTNYGMNWVKQAPGKGLKWMGWINTYTGEPTYADDFK
|*||**||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
QIQLVQSGPELKKPGETVKISCKASGYTFTNYGMNWVKQAPGKGLKWMGWINTYTGEPTYADDFK

GRFAFSLETSASTAYLQINNLKNEDTATYFCVQAERLRRTFDYWGAGTTVTVTS
|||||||||||||||||||||||||||||||**|||| **||*||||||||||
GRFAFSLETSASTAYLQINNLKNEDTATYFCARAERL-YWFDVWGAGTTVTVTS

1A4J Light
ELVMTQTPLSLPVSLGDQASISCRSSQSLVHSNGNTYLHWYLQKPGQSPKLLIYKVSNRFSGVPD
**||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
DVVMTQTPLSLPVSLGDQASISCRSSQSLVHSNGNTYLHWYLQKPGQSPKLLIYKVSNRFSGVPD

RFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHVPPTFGGGTKLEIKR-
|||||||||||||||||||||||||||||||||||||||||||||||*|
RFSRVEAEDLGVYFCSQSTHVPPTFGGGTKLEINRASGSGSGTDFTLKI

1BLN Heavy
EVILVESGGGLVKPGGSLKLSCAASGFTFSSYTMSWVRQTPEKRLEWVATISSGGGNTYYPDSVK
||*||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
EVMLVESGGGLVKPGGSLKLSCAASGFTFSSYTMSWVRQTPEKRLEWVATISSGGGNTYYPDSVK

GRFTISRDNAKNNLYLQMSSLRSEDTALYYCARYYRYEAWFASWGQGTLVTVTS
|||||||||||||||||||||||||||||||||||||||||*||||||||||
GRFTISRDNAKNNLYLQMSSLRSEDTALYYCARYYRYEAWFAYWGQGTLVTVTS
```

```
1BLN Light

DVLMTQTPVSLSVSLGDQASISCRSSQSIVHSTGNTYLEWYLQKPGQSPKLLIYKISNRFSGVPD

||||||||*||*||||||||||||||||||*||||||||||||||||||||*|||||||||

DVLMTQTPLSLPVSLGDQASISCRSSQSIVHSNGNTYLEWYLQKPGQSPKLLIYKVSNRFSGVPD


RFSGSGSGTDFTLKISRVEAEDLGVYYCFQASHAPRTFGGGTKLEIKR-

||||||||||||||||||||||||||||||*||*|*||||||||||||

RFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVPWTFGGGTKLEIKRA


1IGF Heavy

EVQLVESGGDLVKPGGSLKLSCAASGFTFSRCAMSWVRQTPEKRLEWVAGISSGGSYTFYPDTVK

|||||||||*||||||||||||||||**||||||||||||||||*||||||||*|||*||

EVQLVESGGGLVKPGGSLKLSCAASGFTFSSYAMSWVRQTPEKRLEWVATISSGGSYTYYPDSVK


GRFIISRNNARNTLSLQMSSLRSEDTAIYYCTRYSSDPFYFDYWGQGTTLTVS

|||*|||*||*|||*||||||||||||*|||*||||||||||||||||||||

GRFTISRDNAKNTLYLQMSSLRSEDTAMYYCARYSSDPFYFDYWGQGTTLTVS


1IGF Light

DVLMTQTPLSLPVSLGDQASISCRSNQTILLSDGDTYLEWYLQKPGQSPKLLIYKVSNRFSGVPD

||||||||||||||||||||||||||*|*|**|*|*||||||||||||||||||||||||||||

DVLMTQTPLSLPVSLGDQASISCRSSQSIVHSNGNTYLEWYLQKPGQSPKLLIYKVSNRFSGVPD


RFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVPPTFGGGTKLEIKR-

||||||||||||||||||||||||||||||||||||||||||||||||||

RFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVPPTFGGGTKLEIKRA


1RUR Heavy

EVQLEESGPELVRPGTSVKISCKASGYTFTNYWLGWVKQRPGHGFEWIGDIYPGGVYTTNNEK
```

```
*|||**||*|||||||||*||||*||||||||*|||||||||*|||||||||*||**|||
QVQLQQSGAELVRPGTSVKMSCKAAGYTFTNYWIGWVKQRPGHGLEWIGDIYPGGGYTNYNEK

FRGKAILTADTSSSTAYMQLSSLTSEDSAVYFCARAGGYYTGGDYWGQGTSVTVS
|*|||*|||||||||||||||||||||||*|*||||||||**||||||||*|||
FKGKATLTADTSSSTAYMQLSSLTSEDSAIYYCARAGGYYTYFDYWGQGTSLTVS
```

1RUR Light

```
DIVLTQAAFSNPVTLGASASISCRSSKSLLNSNGIIHMYWYLQKPGQSPQLLIYQMSKLASGAPD
|||*||||||||||||*||||||||||||*||||***||||||||||||||||||*||||*||
DIVMTQAAFSNPVTLGTSASISCRSSKSLLHSNGITYLYWYLQKPGQSPQLLIYQMSNLASGVPD

RFSGSGSGTDFTLRISRVEAEDVGVYYCAQNLELPYTFGGGTKLEIKR-
|||*|||||||||||||||||||||||||||||||||||||||||||
RFSSSGSGTDFTLRISRVEAEDVGVYYCAQNLELPYTFGGGTKLEIKRA
```

1RZ7 Heavy

```
EVQLVQSGAEVKKPGATVKISCKASGYTFSDFYMYWVRQAPGKGLEWMGLIDPEDADTMYAEKFR
||||||||||||||||||||||*|||||*|*||*||*||||||||||||*||||**|*|||||*
EVQLVQSGAEVKKPGATVKISCKVSGYTFTDYYMHWVQQAPGKGLEWMGLVDPEDGETIYAEKFQ

GRVTITADTSTDTGYLELSSLRSEDTAVYYCAADPWELNAFNVWGQGTLVSVSS
||||||||||||*|*||||||||||||||*||||||*|**|||||||*| |
GRVTITADTSTDTAYMELSSLRSEDTAVYYCATDPWELNYFDYWGQGTLVTV-S
```

1RZ7 Light

```
IQMTQSPSSVSASVGDRVTITCRASQDISTWLAWYQQKPGKAPKLLIYAASTLQSGVPSRFSGSG
||||||||||||||||||||||||||*||*||||||||||||||||||*||||||||||||||
IQMTQSPSSVSASVGDRVTITCRASQGISSWLAWYQQKPGKAPKLLIYAASSLQSGVPSRFSGSG
```

```
SGTDFSLTINSLQPEDFATYYCQQANSF-FTFGGGTKVEIKRT
|||||*|||*|||||||||||||||| *|||||||||||||
SGTDFTLTISSLQPEDFATYYCQQANSFPLTFGGGTKVEIKRT
```

1T2Q Heavy

```
EVQLLEESGPGLVQPSQSLSITCTVSGFSLTSYGVHWVRQSPGKGLEWLGVIWSGGSTDYNAAFI
|***|**|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
EQVQLKQSGPGLVQPSQSLSITCTVSGFSLTSYGVHWVRQSPGKGLEWLGVIWSGGSTDYNAAFI

SRLSISKDNSKSQVFFKMNSLQADDTAIYYCARNRGYSYAMDSWGQGTSVTVS
||||||||||||||||||||||||||||||||||*|||*||||||||||
SRLSISKDNSKSQVFFKMNSLQADDTAIYYCARNRGYYYAMDYWGQGTSVTVS
```

1T2Q Light

```
ELVMTQSPLSLPVSLGDQASISCRSSQSLVHSSGNTYLHWYLQKPGQSPKLLIYKVSNRFSGVPD
***|||*|||||||||||||||||||*|||*|||||*|||||||||||||||||||||||||||
DVLMTQTPLSLPVSLGDQASISCRSSQSIVHSNGNTYLEWYLQKPGQSPKLLIYKVSNRFSGVPD

RFSGSGSGTDFTLTISRVEAEDLGVYYCFQGSHVPLTFGAGTKLELKR-
||||||||||||*||||||||||||||||||||||||||||||||||
RFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVPLTFGAGTKLELKRA
```

2AGJ Heavy

```
VTLKESGPTLVKPTQTLTLTCTFSGFSLTTTGEGVGWIRQPPGKALEFLAFIYWNDAKRYNPSLQ
*|||||||||||||||||||||||||*|*|*|||||||||||||*||*|||||*|||*|||*
ITLKESGPTLVKPTQTLTLTCTFSGFSLSTSGVGVGWIRQPPGKALEWLALIYWNDDKRYSPSLK

SRLTITKDASKKQVVLTLTNLDPVDTATYYCARTSGWDIEFEYWGQGTLVTVS
```

```
||||||||*||*|||||*||*||||||||||**|||||*|*||||||||||
SRLTITKDTSKNQVVLTMTNMDPVDTATYYCAHRSGWDIYFDYWGQGTLVTVS

2AGJ Light
EIVLTQSPGTLSLSPGERATLSCRASETVSNDKVAWYQQKPGQAPRLLIYGASSRATGIPDRFSG
||||||||||||||||||||||||||**||****||||||||||||||||||||||||||||||
EIVLTQSPGTLSLSPGERATLSCRASQSVSSSYLAWYQQKPGQAPRLLIYGASSRATGIPDRFSG

SGSGTDFTLSISGLEPEDFVVYYCQQYASSPRTFGQGTKVEIKRT
|||||||||*||*||||||*|||||||*|||*||||||||||||*
SGSGTDFTLTISRLEPEDFAVYYCQQYGSSPWTFGQGTKVEIKRL

3KYM Heavy
EVQLLESGGGLVQPGGSLRLSCAASGFTFSIYPMFWVRQAPGKGLEWVSWIGPSGGITKYADSVK
||||||||||||||||||||||||||||||*|*|*||||||||||||||*|**|||*|*||||||
EVQLLESGGGLVQPGGSLRLSCAASGFTFSSYAMSWVRQAPGKGLEWVSAISGSGGSTYYADSVK

GRFTISRDNSKNTLYLQMNSLRAEDTATYYCAREGHNDWYFDLWGRGTLVTVS
|||||||||||||||||||||||||*||||*||||*||||||||||||||||
GRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAKEGHNYWYFDLWGRGTLVTVS

3KYM Light
DIQMTQSPGTLSLSPGERATLSCRASQSVSSYLAWYQQKPGQAPRLLIYDASNRATGIPARFSGS
*|*|||||*|||*||||||||||||||||*||||||||||||||||*||*||||||||||||
EIVMTQSPATLSVSPGERATLSCRASQSVSSNLAWYQQKPGQAPRLLIYGASTRATGIPARFSGS

GSGTEFTLTISSLQSEDFAVYYCQQYDKWPLTFGGGTKVEIK
|||||||||||||||||||||||||**||||||||||||||
GSGTEFTLTISSLQSEDFAVYYCQQYNNWPLTFGGGTKVEIK
```

```
3QRG Heavy

ITLKESGPTLVKPTQTLTLTCTFSGFSLSTSGMGVSWIRQPPGKALEWLAHIYWDDDKRYNPSLK

||||||||||||||||||||||||||||||*||*|||||||||||||*|||||||||*||||

ITLKESGPTLVKPTQTLTLTCTFSGFSLSTSGVGVGWIRQPPGKALEWLALIYWDDDKRYSPSLK


SRLTITKDTSKNQVVLTMTNMDPVDTATYYCARLYGFTYGFAYWGQGTLVTVS

||||||||||||||||||||||||||||||**|||||*|*||||||||||||

SRLTITKDTSKNQVVLTMTNMDPVDTATYYCAHRYGFTYYFDYWGQGTLVTVS


3QRG Light

DIVMTQSPDSLAVSLGERATINCRASQSVDY--NGISYMHWYQQKPGQPPKLLIYAASNPESGVPD

|||||||||||||||||||||||**||||*|  |***|**||||||||||||||||*||**||||||

DIVMTQSPDSLAVSLGERATINCKSSQSVLYSSNNKNYLAWYQQKPGQPPKLLIYWASTRESGVPD


RFSGSGSGTDFTLTISSLQAEDVAVYYCQQIIEDPWTFGQGTKVEIKR-

||||||||||||||||||||||||||||||****||||||||||||||

RFSGSGSGTDFTLTISSLQAEDVAVYYCQQYYSTPWTFGQGTKVEIKRT


4YGV Heavy

QVQLVQSGAEVKKPGASVKVSCKASGYTFTDYYMHWVRQAPGQGLEWMGETNPRNGGTTYNEKFK

||||||||||||||||||||||||||||||*||||||||||||||||||**||**|*|*|**||*

QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYYMHWVRQAPGQGLEWMGIINPSGGSTSYAQKFQ


GKATMTRDTSTSTAYMELSSLRSEDTAVYYCTIGTSGYDYFDYWGQGTLVTVS

|**||||||||||*||||||||||||||**||||||||||||||||||||||

GRVTMTRDTSTSTVYMELSSLRSEDTAVYYCARGTSGYDYFDYWGQGTLVTVS


4YGV Light

DIVMTQTPLSLSVTPGQPASISCRSSQSIVHSDGNIYLEWYLQKPGQSPKLLIYKVSYRFSGVPD
```

```
||||||||||||||||||||||*||||**||||**||*|||||||||*||||*||*|||||||
DIVMTQTPLSLSVTPGQPASISCKSSQSLLHSDGKTYLYWYLQKPGQSPQLLIYEVSSRFSGVPD

RFSGSGSGTDFTLKISRVEAEDVGVYYCFQASHVPYTFGQGTKLEIKR-
||||||||||||||||||||||||||*|**|*|||||||||||||||
RFSGSGSGTDFTLKISRVEAEDVGVYYCMQGIHLPYTFGQGTKLEIKRT
```

4.A.4 Comparison of flexibility calculations across ensemble generation methods

In this work, I have considered multiple ensemble generation methods in conjunction with FIRST-PG analysis to determine the flexibility of CDR-H3 loops. Of all methods used in this paper to generate structural ensembles, only MD simulations have previously been coupled with flexibility analysis. Rosetta FastRelax, KIC, and RosettaAntibody have not been used previously for flexibility analysis. MD simulations permit for fluctuations between low-energy states and variations in hydrogen bonding networks, effectively capturing the "flickering" nature of hydrogen bonds. The Rosetta-based methods consider hydrogen-bonding energy, but do not involve dynamic motion in the same way as MD simulations. The Rosetta FastRelax protocol generates ensembles representative of the local energy minimum through side-chain repacking and gradient-based energy minimization, so flexibility analysis of these ensembles should be comparable to flexibility analysis of crystal structures. Rosetta KIC on the other hand generates ensembles of low-energy CDR-H3 loop conformations by *de novo* modeling of the CDR-H3 loop. RosettaAntibody generates ensembles of low-energy antibody conformations, building on KIC motions through additional VH–VL docking. I compared FIRST-PG calculations on Rosetta FastRelax, RosettaAntibody, and MD ensembles, for three well-studied antibodies, excluding KIC ensembles from analysis because they are effectively superseded by RosettaAntibody ensembles. Qualitatively, the FIRST-PG results agree for all methods for 48G7 and the anti-fluorescein antibody. The Rosetta FastRelax results differ from the other methods

for the anti-influenza, with the naïve showing significantly more rigidity. This is most likely due to the difference in quality between the crystal structures. Quantitatively, the ΔAUC values for RosettaAntibody and MD ensembles for all three antibody pairs compare well (Supplemental Table 4.A.1). Additionally, I compared only RosettaAntibody and MD ensembles for three naïve-reverted/mature antibody pairs, where I found that the ΔAUC values roughly agree for two out of three antibody pairs. Taken together, these results indicate that flexibility analyses on RosettaAntibody homology model ensembles are similar to analyses on MD ensembles.

4.A.5 Supplemental tables

Table 4.A.1: Changes in the rigidity of the 48G7 antibody CDR-H3 loop according to several methods. Unbound is denoted by (U) and bound is denoted by (B). A positive number indicates an increase in rigidity upon affinity maturation. Changes for B-factors are calculated as the difference in the average CDR-H3 loop B-factor between the naïve and mature crystal structure: $\Delta B = B_{naive} - B_{mature} \pm \sqrt{s^2_{naive} + s^2_{mature}}$. Changes in FIRST-PG are calculated as the percent change between the AUC of the CDR-H3 melting curve for naïve and mature antibodies: $\Delta AUC = 100 \times \frac{AUC_{mature} - AUC_{naive}}{AUC_{naive}}$. Finally, changes in MD RMSD or RMSF are calculated as the difference in average CDR-H3 loop RMSF or RMSD between the MD simulations of the naïve and mature antibodies: $\Delta R = R_{naive} - R_{mature} \pm \sqrt{s^2_{naive} + s^2_{mature}}$. Only bound crystal structures were available for the 4-4-20 antibody, but Relax, KIC, RA and MD simulations were run without antigen. Only an unbound naïve and bound mature crystal structures were available for the anti-influenza antibody, but Relax, KIC, RA and MD simulations were run without antigen.

Antibody	ΔB-Factor	ΔRelax AUC	ΔKIC AUC	ΔRA AUC	ΔMD RMSD	ΔMD RMSF	ΔMD AUC
48G7 (U)	2.14 ± 0.62	3.9	13.0	0.2	-1.04 ± 0.66	2.14 ± 0.62	-1.3
48G7 (B)	1.21 ± 0.89	-6.2	-8.9				
4-4-20 (U)				-6.2	0.85 ± 0.53	-0.35 ± 0.36	-4.1
4-4-20 (B)	0.46 ± 0.77	-8.4	2.8				
Influenza (U)				6.1	2.25 ± 1.33	0.44 ± 1.14	9.1
Influenza (B)	1.85 ± 1.64	3.9	-15.2	1.7			

4.A.6 Supplemental figures

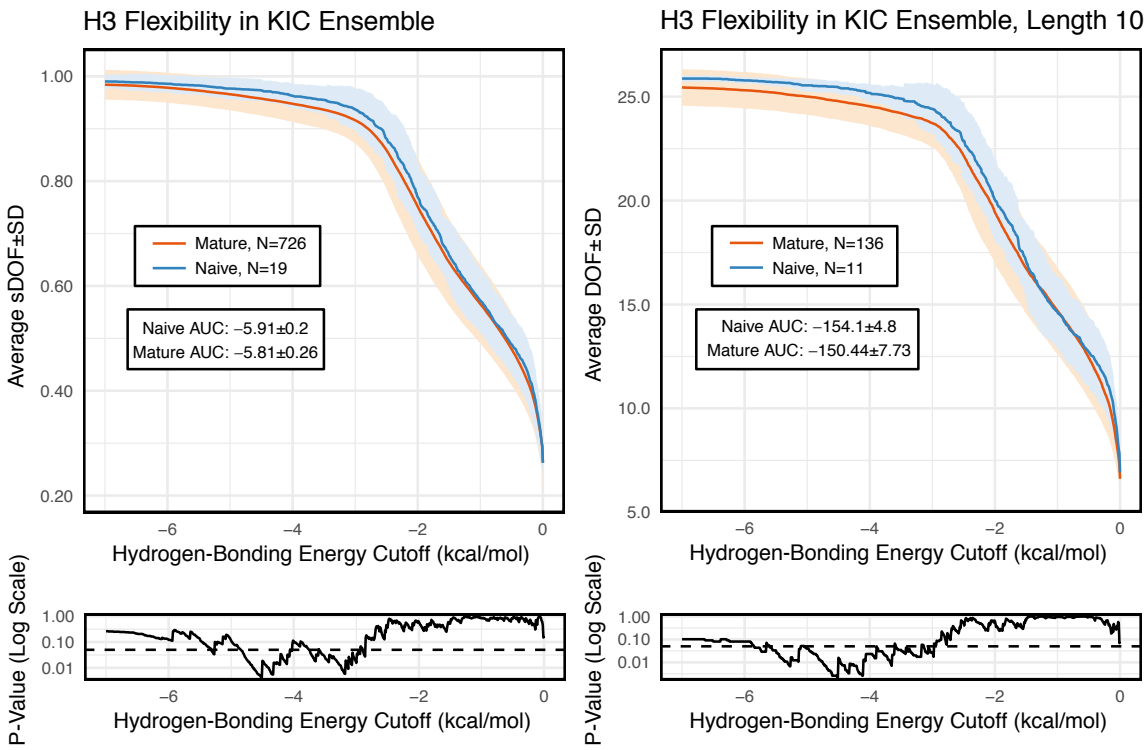

Figure 4.A.1: FIRST-PG analysis of KIC ensembles of the crystallographic antibody set, with naïve antibody data shown in blue and mature antibody data shown in orange and standard error of the mean shown in a lighter shade of the respective color. Subplots, below each main plot, show the p-value computed by a KS comparison of the naïve and mature DOF distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same and a dashed line indicating a p-value of 0.05. (Left) When comparing DOFs scaled to a theoretical maximum as a function of hydrogen-bonding energy cutoff for the entire set, the values are similar for both naïve (AUC = -5.9 ± 0.2) and mature (AUC = -5.8 ± 0.3) antibodies. (Right) Comparison of DOFs for a single length without scaling reveals naïve antibodies to possess a slightly higher DOF value than mature antibodies at the same hydrogen-bonding energy cutoff. AUCs however are within a standard deviation, compare naïve at -154.1 ± 4.8 and mature at -150.4 ± 7.7.

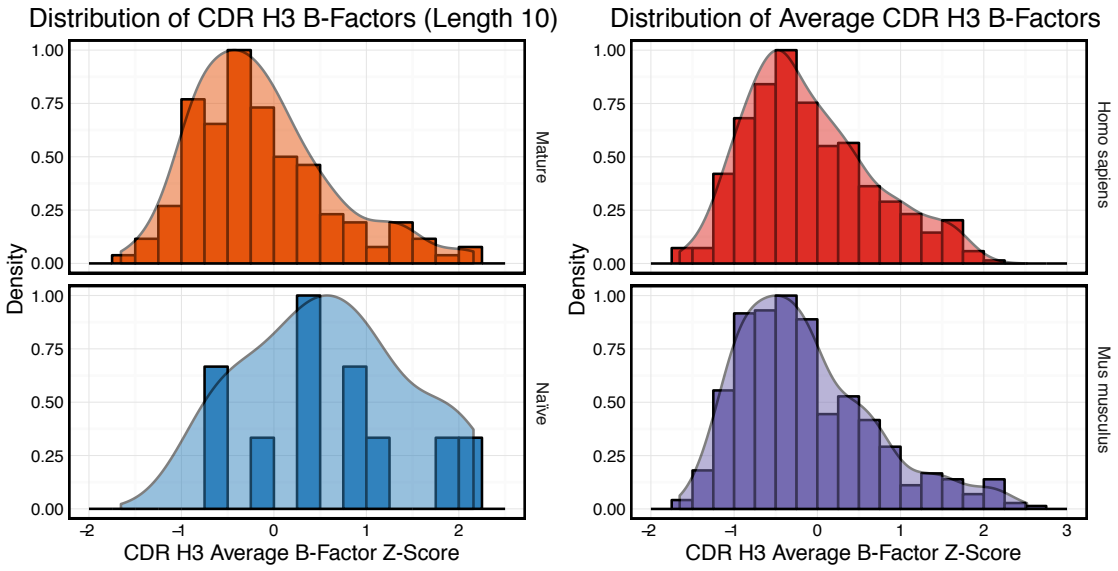

Figure 4.A.2: Average CDR-H3 loop B-factor z-score for antibodies with loops of length 10 split by number of mutations (left) and all antibodies split by heavy-chain species (right). Mature antibodies have at least one mutation. Comparing the difference in mature versus naïve means for length 10 CDR-H3 loops only to a randomized test (as described in the methods) shows only 2.9% of random permutations have an equal or greater difference. A two-sample KS test yields a p-value of 0.0135 and D of 0.4949, so these distributions appear to be on the threshold of significance. However, that is obviated when bound structures are excluded from analysis, resulting in 4.8% of random permutations having an equal or greater difference in means than the observed and a KS-test p-value of 0.0989 (with D of 0.3989). It is difficult to quantify if there is a difference in the length 10 set due to low counts (only 11 naïve antibodies), whereas there is no visible difference between the human and mouse antibodies (21.3% of random permutations have an equal or greater difference and the KS-test p-value is 0.6654 [with a D of 0.0748]).

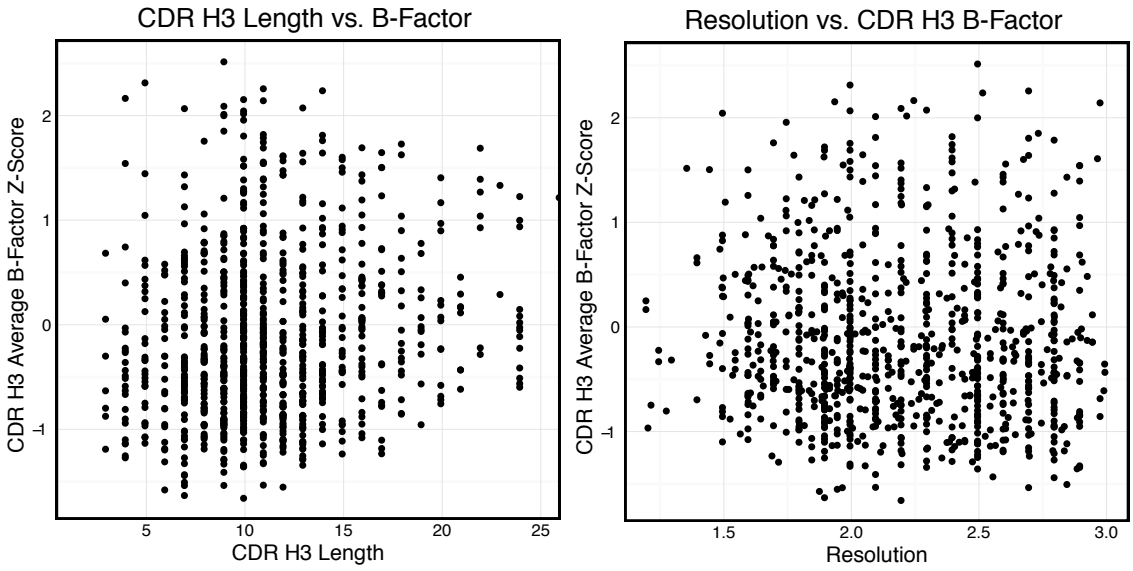

Figure 4.A.3: Average CDR-H3 loop B-factor z-score compared with either loop length (left) or crystal structure resolution (right). There is not an obvious dependence of CDR-H3 loop B-factor z-score on either.

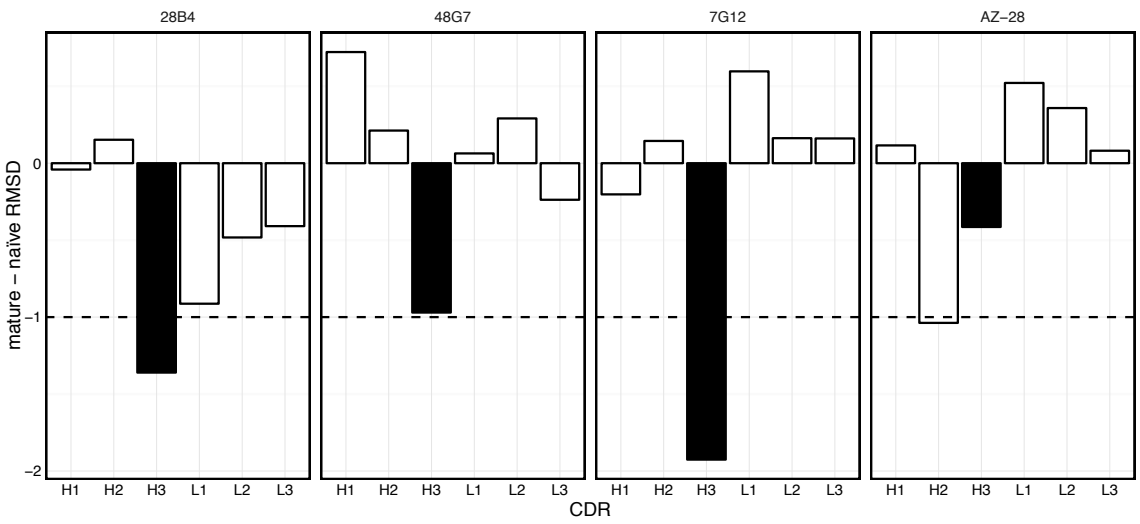

Figure 4.A.4: Difference in CDR loop motions upon antigen binding between naïve and mature antibodies for four catalytic antibodies. Loop RMSDs (in angstroms) were calculated from the difference in CÎś atom positions after alignment of the corresponding (heavy or light) framework CÎś atoms. The CDR-H3 loops is highlighted in black. The dashed line indicates 1 Å. A more negative value here indicates less motion upon binding in the mature antibody. The effects of affinity maturation on CDR-H3 loop motion in crystal structures are not always significant, with only 2/4 showing motion reduction greater than an angstrom.

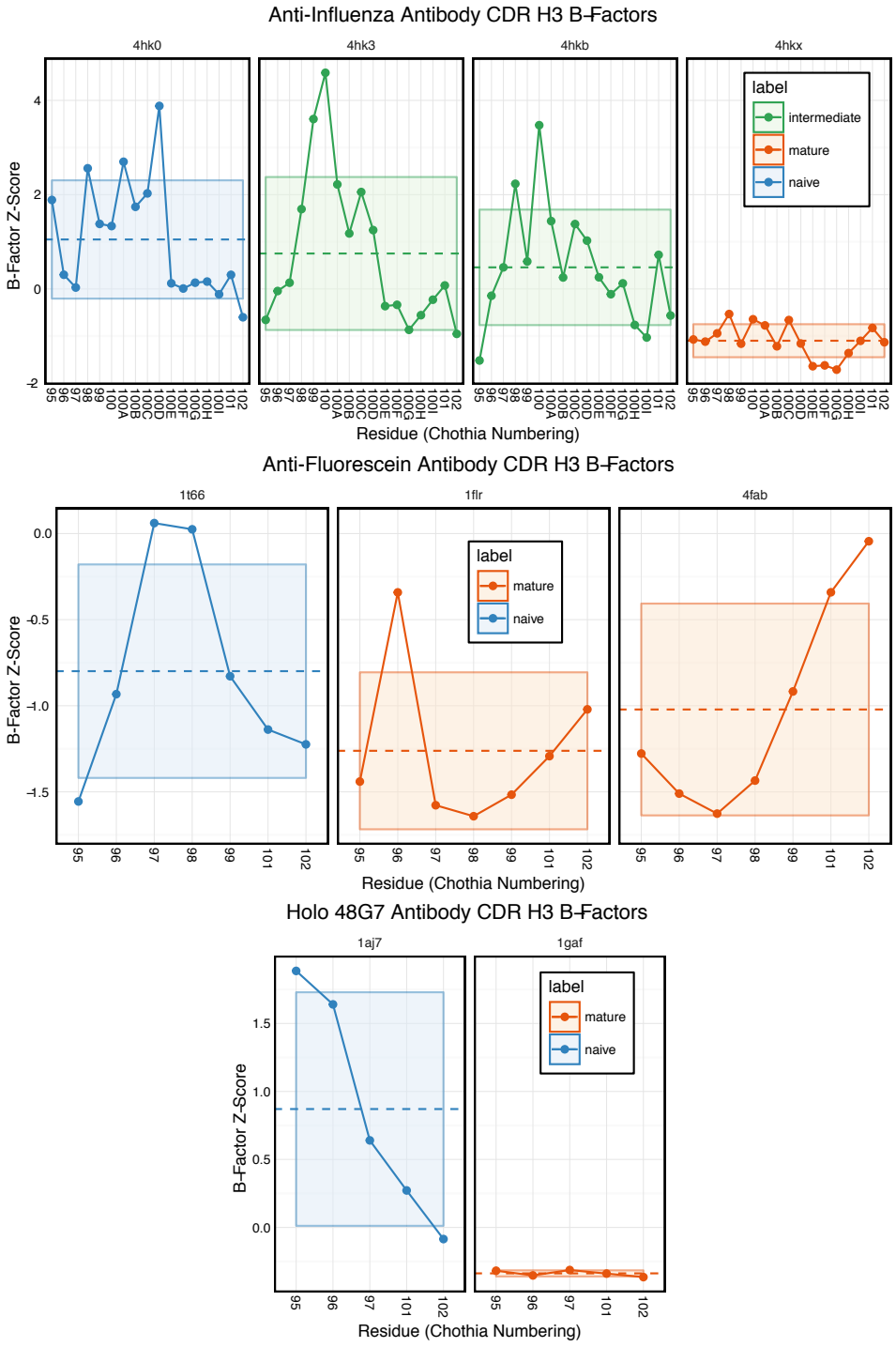

Figure 4.A.5: Caption follows on the next page.

Figure 4.A.5: (Previous page.) CDR-H3 loop B-factor z-scores for three previously studied antibodies, with PDB IDs shown above each plot. B-factor z-scores were calculated with respect to the Fv region and for Cα atoms only. The anti-influenza antibodies have vary in resolution from 2.5 Å for the naïve and mature to 3.0 Å (4HK3) and 3.6 Å (4HKB) for the intermediates. Additionally, the mature anti-influenza antibody has antigen bound affecting the CDR-H3 loop B-factors. One can see that affinity maturation does not always lead to a reduction in CDR-H3 loop B-factor z-scores.

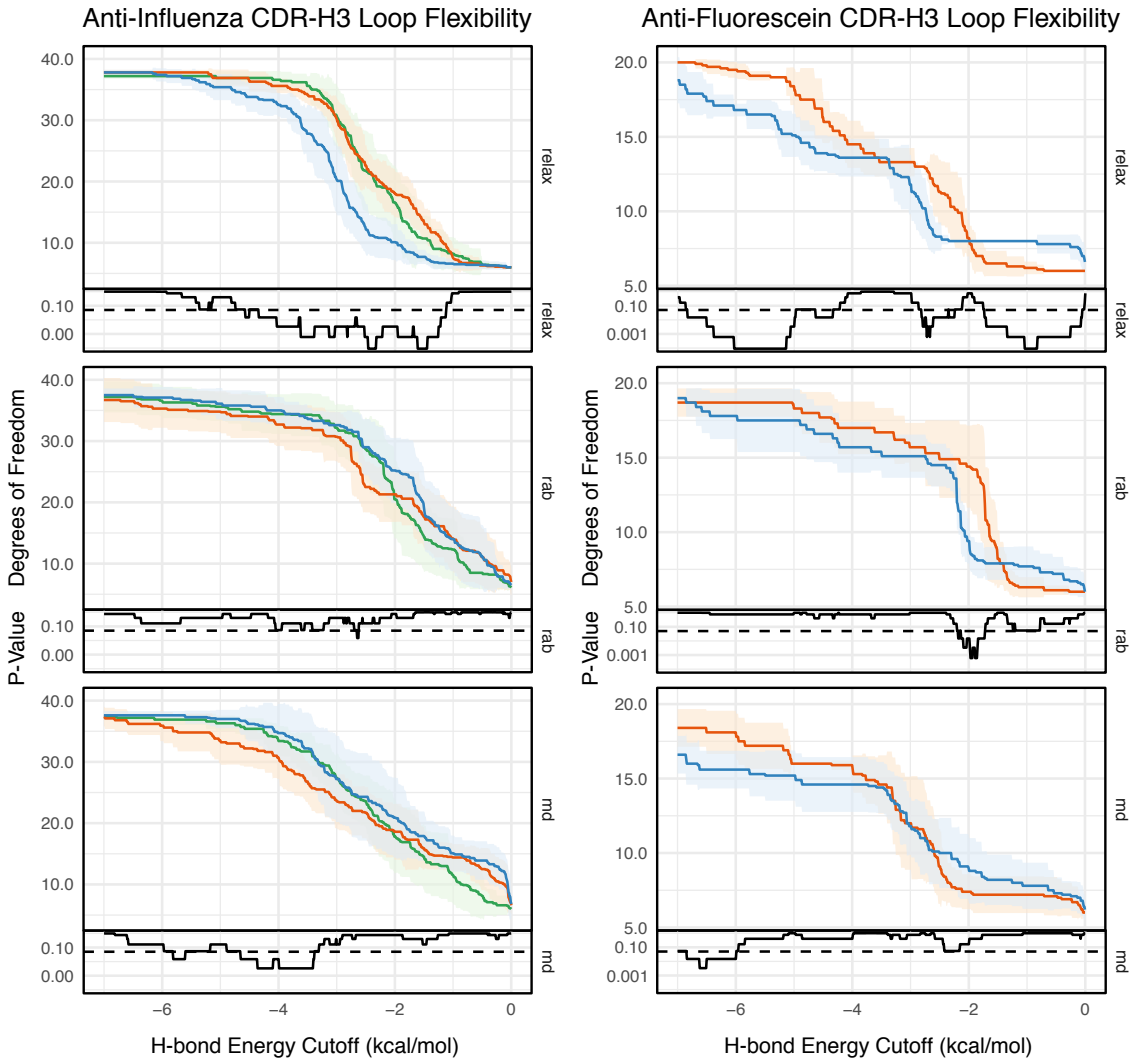

Figure 4.A.6: FIRST-PG analysis of two previously studied antibodies with MD simulations (labelled MD), RosettaAntibody (labelled RAB), and Rosetta FastRelax (labelled relax) used to generate structural ensembles. Naïve antibodies are colored blue and mature antibodies are colored red, while an "intermediate" (4HKB) influenza antibody is shown in green. Subplots, below each main plot, show the p-value computed by a KS comparison of the naïve and mature DOF distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same and a dashed line indicating a p-value of 0.05. Again, the effects of affinity maturation are not obvious.

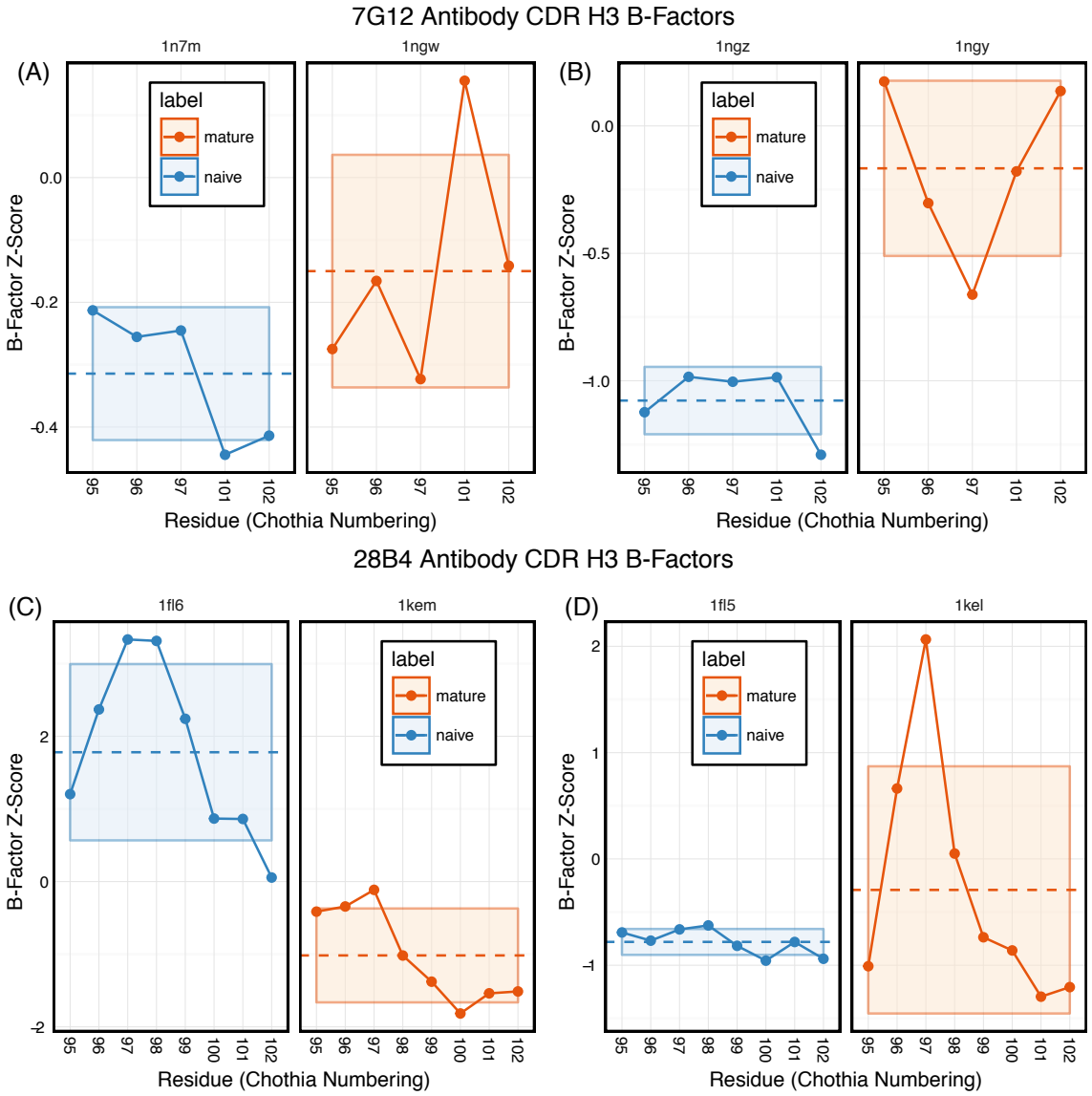

Figure 4.A.7: CDR-H3 B-factor z-scores for antigen-bound and free crystal structures of catalytic antibodies 7G12 and 28B4. The 7G12 antibody has higher z-scores for the mature than the naïve antibody for both the (A) unbound and (B) bound structures, indicating a gain in flexibility upon maturation. The 28B4 antibody shows a loss of flexibility upon maturation for the unbound structure comparison (C), but no change in the bound structure comparison (D).

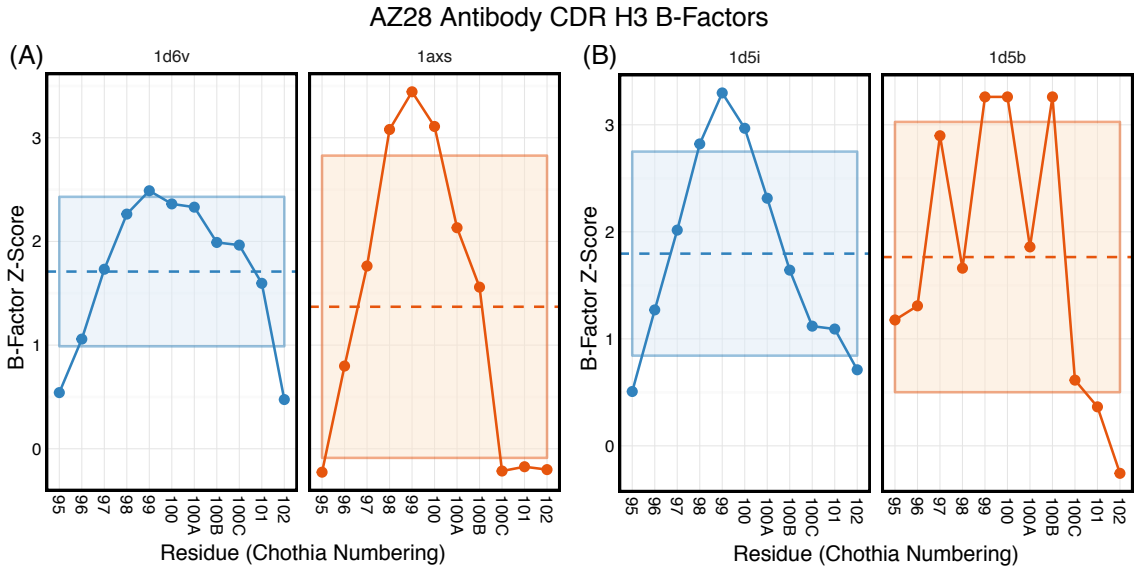

Figure 4.A.8: CDR-H3 loop B-factor z-scores for antigen-bound and free crystal structures of the catalytic antibody AZ-28 reveal no significant difference between the naïve and mature antibodies.

References

1. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302,** 575–581 (1983).
2. Di Noia, J. M. & Neuberger, M. S. Molecular Mechanisms of Antibody Somatic Hypermutation. *Annual Review of Biochemistry* **76,** 1–22 (2007).
3. De los Rios, M., Criscitiello, M. F. & Smider, V. V. *Structural and genetic diversity in antibody repertoires from diverse species* 2015.
4. Clark, L. A., Ganesan, S, Papp, S & van Vlijmen, H. W. Trends in antibody sequence changes during the somatic hypermutation process. *J Immunol* **177,** 333–340 (2006).
5. Burkovitz, A., Sela-Culang, I. & Ofran, Y. Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. en. *FEBS Journal* **281,** 306–319 (2014).
6. Chothia, C. & Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* **196,** 901–917 (1987).
7. Chothia, C. *et al.* Conformations of immunoglobulin hypervariable regions. *Nature* **342,** 877–883 (1989).
8. Al-Lazikani, B., Lesk, A. M. & Chothia, C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* **273,** 927–948 (1997).
9. Kuroda, D., Shirai, H., Kobori, M. & Nakamura, H. Systematic classification of CDR-L3 in antibodies: implications of the light chain subtypes and the VL-VH interface. *Proteins* **75,** 139–146 (2009).
10. North, B., Lehmann, A. & Dunbrack, R. L. A new clustering of antibody CDR loop conformations. *Journal of Molecular Biology* **406,** 228–256 (2011).
11. Morea, V., Tramontano, A., Rustici, M., Chothia, C. & Lesk, A. M. Conformations of the third hypervariable region in the VH domain of immunoglobulins 1 1Edited by I. A. Wilson. *Journal of Molecular Biology* **275,** 269–294 (1998).
12. Kuroda, D., Shirai, H., Kobori, M. & Nakamura, H. Structural classification of CDR-H3 revisited: A lesson in antibody modeling. *Proteins: Structure, Function and Genetics* **73,** 608–620 (2008).
13. Weitzner, B. D., Dunbrack, R. L. & Gray, J. J. The origin of CDR H3 structural diversity. *Structure* **23,** 302–11 (2015).
14. Tsuchiya, Y & Mizuguchi, K. The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops. *Protein Sci* **25,** 815–825 (2016).

15. Wedemayer, G. J., Patten, P. A., Wang, L. H., Schultz, P. G. & Stevens, R. C. Structural insights into the evolution of an antibody combining site. *Science* **276,** 1665–1669 (1997).

16. Mundorff, E. C. *et al.* Conformational effects in biological catalysis: An antibody-catalyzed oxy-Cope rearrangement. *Biochemistry* **39,** 627–632 (2000).

17. Yin, J. *et al.* A comparative analysis of the immunological evolution of antibody 28B4. *Biochemistry* **40,** 10764–10773 (2001).

18. Yin, J., Beuscher, A. E., Andryski, S. E., Stevens, R. C. & Schultz, P. G. Structural plasticity and the evolution of antibody affinity and specificity. *Journal of molecular biology* **330,** 651–656 (2003).

19. Li, Y., Li, H., Yang, F., Smith-Gill, S. J. & Mariuzza, R. A. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nature Structural & Molecular Biology* **10,** 482–488 (2003).

20. Manivel, V., Sahoo, N. C., Salunke, D. M. & Rao, K. V. Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity* **13,** 611–620 (2000).

21. Schmidt, A. G. *et al.* Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proc Natl Acad Sci U S A* **110,** 264–9 (2013).

22. Thielges, M. C., Zimmermann, J., Yu, W., Oda, M. & Romesberg, F. E. Exploring the energy landscape of antibody-antigen complexes: Protein dynamics, flexibility, and molecular recognition. *Biochemistry* **47,** 7237–7247 (2008).

23. Adhikary, R., Yu, W., Oda, M., Zimmermann, J. & Romesberg, F. E. Protein dynamics and the diversity of an antibody response. *J Biol Chem* **287,** 27139–27147 (2012).

24. Adhikary, R. *et al.* Adaptive mutations alter antibody structure and dynamics during affinity maturation. *Biochemistry* **54,** 2085–2093 (2015).

25. Jimenez, R., Salazar, G., Baldridge, K. K. & Romesberg, F. E. Flexibility and molecular recognition in the immune system. *Proc Natl Acad Sci U S A* **100,** 92–97 (2003).

26. Jimenez, R., Salazar, G., Yin, J., Joo, T. & Romesberg, F. E. Protein dynamics and the immunological evolution of molecular recognition. *Proc Natl Acad Sci U S A* **101,** 3803–3808 (2004).

27. Zimmermann, J. J. *et al.* Antibody evolution constrains conformational heterogeneity by tailoring protein dynamics. *Proc Natl Acad Sci U S A* **103,** 13722–13727 (2006).

28. Davenport, T. M. *et al.* Somatic Hypermutation-Induced Changes in the Structure and Dynamics of HIV-1 Broadly Neutralizing Antibodies. *Structure* **24,** 1346–1357 (2016).

29. Chong, L. T., Duan, Y, Wang, L, Massova, I & Kollman, P. A. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 14330–14335 (1999).

30. Thorpe, I. F., Brooks, C. L. & Brooks 3rd, C. L. Molecular evolution of affinity and flexibility in the immune system. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 8821–8826 (2007).

31. Wong, S. E., Sellers, B. D. & Jacobson, M. P. Effects of somatic mutations on CDR loop flexibility during affinity maturation. en. *Proteins: Structure, Function and Bioinformatics* **79,** 821–829 (2011).

32. Li, T. *et al.* Rigidity Emerges during Antibody Evolution in Three Distinct Antibody Systems: Evidence from QSFR Analysis of Fab Fragments. *PLoS Comput Biol* **11** (ed de Groot, B. L.) e1004327 (2015).

33. Di Palma, F. & Tramontano, A. Dynamics behind affinity maturation of an anti-HCMV antibody family influencing antigen binding. *FEBS Letters* **591,** 2936–2950 (2017).

34. Babor, M. & Kortemme, T. Multi-constraint computational design suggests that native sequences of germline antibody H3 loops are nearly optimal for conformational flexibility. *Proteins: Structure, Function, and Bioinformatics* **75,** 846–858 (2009).

35. Willis, J. R. *et al.* Human germline antibody gene segments encode polyspecific antibodies. *PLoS Comput Biol* **9,** e1003045 (2013).

36. DeKosky, B. J. *et al.* Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proceedings of the National Academy of Sciences* **113,** E2636–E2645 (2016).

37. Weitzner, B. D. *et al.* Modeling and docking of antibody structures with Rosetta. *Nature Protocols* **12,** 401–416 (2017).

38. Weitzner, B. D., Kuroda, D., Marze, N., Xu, J. & Gray, J. J. Blind prediction performance of RosettaAntibody 3.0: Grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins: Structure, Function and Bioinformatics* **82,** 1611–1623 (2014).

39. Sljoka, A. *Algorithms in rigidity theory with applications to protein flexibility and mechanical linkages* PhD thesis (York University, 2012).

40. Sljoka, A. & Wilson, D. Probing protein ensemble rigidity and hydrogen-deuterium exchange. *Physical Biology* **10,** 056013 (2013).

41. Kim, H. & Ha, T. Single-molecule nanometry for biological physics. *Reports on progress in physics. Physical Society (Great Britain)* **76,** 016601 (2013).

42. Deng, B. *et al.* Suppressing allostery in epitope mapping experiments using millisecond hydrogen / deuterium exchange mass spectrometry. *mAbs* **9,** 1327–1336 (2017).

43. Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins: Structure, Function and Genetics* **44,** 150–165 (2001).

44. Whiteley, W. *Counting out to the flexibility of molecules* in *Physical Biology* **2** (IOP Publishing, 2005), S116–S126.

45. Dunbar, J. *et al.* SAbDab: The structural antibody database. *Nucleic Acids Research* **42,** D1140–D1146 (2014).

46. Ehrenmann, F., Kaas, Q. & Lefranc, M. P. IMGT/3dstructure-DB and IMGT/domaingapalign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MHcSF. *Nucleic Acids Research* **38,** D301–D307 (2009).

47. Jeliazkov, J. R. *et al.* Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification. *Frontiers in Immunology* **9** (2018).

48. Nivon, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS ONE* **8** (ed Zhang, Y.) e59004 (2013).

49. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods* **6,** 551–552 (2009).

50. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13,** 3031–3048 (2017).

51. Chipot, C. *et al.* Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26,** 1781–1802 (2005).

52. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14,** 71–73 (2016).

53. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98,** 10089–10092 (1993).

54. Li, T *et al.* Redistribution of flexibility in stabilizing antibody fragment mutants follows Le Chatelier's principle. *PLoS One* **9,** e92870 (2014).

55. Feig, M., Karanicolas, J. & Brooks, C. L. *MMTSB Tool Set: Enhanced sampling and multiscale modeling methods for applications in structural biology* in *Journal of Molecular Graphics and Modelling* **22** (Elsevier, 2004), 377–395.

56. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry* **32,** 2319–2327 (2011).

57. Mamonova, T., Hespenheide, B., Straub, R., Thorpe, M. F. & Kurnikova, M. Protein flexibility using constraints from molecular dynamics simulations. *Physical Biology* **2,** S137–S147 (2005).

58. Kuroda, D. & Gray, J. J. Pushing the backbone in protein-protein docking. *Structure* **24,** 1821–1829 (2016).

59. Ó Conchúir, S. *et al.* A Web resource for standardized benchmark datasets, metrics, and rosetta protocols for macromolecular modeling and design. *PLoS ONE* **10** (ed Zhang, Y.) e0130433 (2015).

60. Demirel, M. C. & Lesk, A. M. Molecular forces in antibody maturation. *Phys Rev Lett* **95,** 208106 (2005).

61. Murphy, K., Weaver, C. & Mowat, A. *Janeway's Immunobiology* 9th Editio, 1–907 (2017).

62. Boder, E. T., Midelfort, K. S. & Wittrup, K. D. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc Natl Acad Sci U S A* **97,** 10701–10705 (2000).

63. Midelfort, K. S. *et al.* Substantial Energetic Improvement with Minimal Structural Perturbation in a High Affinity Mutant Antibody. *Journal of Molecular Biology* **343,** 685–701 (2004).

64. Fukunishi, H, Shimada, J & Shiraishi, K. Antigen-antibody interactions and structural flexibility of a femtomolar-affinity antibody. *Biochemistry* **51,** 2597–2605 (2012).

65. Xu, H. *et al.* Key mutations stabilize antigen-binding conformation during affinity maturation of a broadly neutralizing influenza antibody lineage. *Proteins* **83,** 771–780 (2015).

66. Eisen, H. N. & Chakraborty, A. K. Evolving concepts of specificity in immune reactions. *Proceedings of the National Academy of Sciences* **107,** 22373–22380 (2010).

67. Alzari, P. M. *et al.* Three-dimensional structure determination of an anti-2-phenyloxazolone antibody : the role of somatic mutation and heavy / light chain pairing in the maturation of an immune response. *The EMBO Journal* **9,** 3807–3814 (1990).

68. Mizutani, R. *et al.* Three-dimensional Structures of the Fab Fragment of Murine N1G9 Antibody from the Primary Immune Response and of its Complex with (4-Hydroxy-3-Nitrophenyl)acetate. *Journal of Molecular Biology* **254,** 208–222 (1995).

69. Yuhasz, S. C., Parry, C., Strand, M. & Amzel, L. M. Structural analysis of affinity maturation: The three-dimensional structures of complexes of an anti-nitrophenol antibody. *Molecular Immunology* **32,** 1143–1155 (1995).

70. Mishra, A. K. & Mariuzza, R. A. *Insights into the structural basis of antibody affinity maturation from next-generation sequencing* 2018.

71. James, L. C., Roversi, P & Tawfik, D. S. Antibody multispecificity mediated by conformational diversity. *Science* **299,** 1362–1367 (2003).

72. Wang, W *et al.* Conformational selection and induced fit in specific antibody and antigen recognition: SPE7 as a case study. *J Phys Chem B* **117,** 4912–4923 (2013).

73. Sterner, E., Peach, M. L., Nicklaus, M. C. & Gildersleeve, J. C. Therapeutic Antibodies to Ganglioside GD2 Evolved from Highly Selective Germline Antibodies. *Cell Reports* **20,** 1681–1691 (2017).

Chapter 5

Hfq Structure Prediction

5.1 Overview

The RNA chaperone Hfq is an Sm protein that facilitates base pairing between bacterial small RNAs (sRNAs) and mRNAs involved in stress response and pathogenesis. Hfq possesses an intrinsically disordered C-terminal domain (CTD) that may tune the function of the Sm domain in different organisms. In *Escherichia coli*, the Hfq CTD increases kinetic competition between sRNAs and recycles Hfq from the sRNA–mRNA duplex. Here, *de novo* Rosetta modeling and competitive binding experiments show that the acidic tip of the *E. coli* Hfq CTD transiently binds the basic Sm core residues necessary for RNA annealing. The CTD tip competes against non-specific RNA binding, facilitates dsRNA release, and prevents indiscriminate DNA aggregation, suggesting that this acidic peptide mimics nucleic acid to auto-regulate RNA binding to the Sm ring. The mechanism of CTD auto-inhibition predicts the chaperone function of Hfq in bacterial genera and illuminates how Sm proteins may evolve new functions.

5.2 Introduction

Over the last twenty years, the traditional paradigm that protein sequence gives rise to structure and that in turn dictates function has been challenged by the emergence of intrinsically disordered[i] proteins (IDPs) and regions (IDRs) within ordered (structured) proteins[1]. More specifically, 8–42% of the residues in the human proteome and 7–30% of residues in the bacterial proteome are predicted to be disordered, depending on the prediction method[2]. Many of these residues partially or entirely constitute proteins with significant biological function. To give two examples from a plethora of possibilities: (1) proteins implicated in human neurodegenerative diseases are disordered[3] and, (2) in bacteria, proteins with IDRs regulate transcription and play a central role in stress response[4].

The study of IDPs and IDRs is necessary not only to understand and prevent human disease, but also to delve into fundamental biological processes. However, it is exceptionally challenging to study disordered elements at a molecular level. This is because, unlike well-ordered or structured proteins, IDPs and IDRs do not occupy one low-energy conformation, instead they exist as a heterogenous and dynamic ensemble of conformations. The heterogeneity renders X-ray crystallography, which relies on atoms existing in repeated and regular positions within the crystal lattice, practically useless, unless there is a single state of interest that can be captured. Furthermore, the dynamic nature of IDP ensembles might span numerous timescales rendering it challenging to characterize with a single technique. Sample experimental approaches to studying IDPs and IDRs include nuclear magnetic resonance (NMR)[5], which is low throughput but provides atomic-scale resolution, single-molecule Förster resonance energy transfer (smFRET)[6], and small-angle X-ray scattering (SAXS)[7], both of which can be high throughput but at a lower resolution.

Complimentary to experimental approaches, computational methods can provide atomic-scale resolution in a high-throughput manner. The most accurate of these ap-

[i]The term, intrinsic disorder, implies that it is the protein sequence that gives rise to the lack of protein structure.

proaches are all-atom molecular dynamics simulations with explicit solvent[8]. Accuracy comes at a steep computational cost: at each time step, atom positions are updated according to the potential energy of the system, which is calculated by summing bonded, van der Waals, and electrostatic over all atom pairs within certain distance cutoffs. Thus, all-atom simulations for large systems are not possible for long time-scales. Simulations can be accelerated by using implicit solvent, such that water molecules (which are the plurality of atoms in a simulation) are not included. Further computational time can be saved by using Monte Carlo, rather than Newtonian sampling. In a Monte Carlo simulation, motions can be user-defined (termed move sets), *e.g.* when modeling a peptide one might only allow changes in backbone and side-chain dihedral angles, as most bond lengths and angles are essentially fixed. In scenarios where all-atom simulations are prohibitively expensive in terms of compute time, Monte Carlo approaches have proven useful in modeling IDPs[9,10] and IDRs[11].

Within Rosetta, the Monte Carlo approach to modeling IDRs is FloppyTail[12]. In comparison to other approaches, FloppyTail is fast, but simplistic, using only backbone and side-chain dihedral moves, and has been applied to a limited set of modeling problems, typically asking whether or not a single IDR can adopt a certain conformation[12–14]. While such information is useful, I thought it would be far more informative to extract biophysical properties (*e.g.* the average strength of residue–residue interactions between the ordered and disordered regions of the protein in the ensemble of plausible models). Thus, I sought to expand the utility of FloppyTail.

To do so, I first identified a well-characterized model system: *E. coli* protein host factor for RNA phage Qβ replication (Hfq). *E. coli* Hfq contains an Sm-like domain (residues 7–65) that oligomerizes into a homohexameric ring with two sequence-specific RNA-binding faces. The proximal face of the ring is highly conserved and binds to uridines[15,16] at the 3'-ends of bacterial small non-coding RNA (sRNA). The distal face of Hfq binds to AAN triplet repeats[17,18] found in mRNA leaders[18,19] and certain sRNAs[20,21]. In addition to

these sequence-specific RNA binding sites, arginine-rich basic patches at the rim of the *E. coli* Hfq hexamer interact with the sRNA body[15,22–24] and facilitate annealing with target mRNAs[25,26]. The *E. coli* Hfq Sm domain is flanked by a short, disordered, N-terminal domain (NTD; residues 1–6), which protrudes from the proximal face of the hexamer, and a longer disordered C-terminal domain (CTD; residues 66–102), which extends from the rim[27,28]. While I use "domain" here and throughout this Chapter, it is not technically correct. Traditionally, the word "domain" refers to a folded protein unit, and is inappropriate when discussing intrinsically disordered segments. Unfortunately, in the Hfq literature CTD and NTD are established terms[29].

Recent work by Dr. Andrew Santiago-Frangos and Professor Sarah Woodson (T.C. Jenkins Department of Biophysics, Johns Hopkins University) showed that the intrinsically disordered CTD of Hfq was involved in RNA displacement from the rim and proximal face of the protein[30]. Working with Dr. Santiago-Frangos, I applied the updated FloppyTail algorithm on Hfq, elucidating the molecular mechanism by which the CTD regulates Hfq activity and improving IDR modeling in Rosetta. In particular, I enabled simultaneous modeling of multiple disorder regions, I identified a criterion for convergence, defining a reasonable simulation length, and I developed an ensemble-based analysis method to extract biophysically relevant properties from models. Simulations of Hfq produced accurate predictions for the energetic effects of mutations on CTD–rim interactions and molecular models congruent with experimental data. Confident in the validity of the approach, I computationally characterized Hfq proteins from other bacterial species, demonstrating a correlation between certain interactions and the activity of these proteins *in vivo*. Finally, a recently-determined crystal structure of the *Caulobacter crescentus* Hfq protein revealed that low-scoring models accurately recapitulate a crystallizable IDR interaction.

5.3 Methods

5.3.1 Structure preparation

In this study, I modeled six Hfq proteins, starting from crystal structures: *E. coli* (1HK9), *P. aeruginosa* (1U1S), *L. monocytogenes* (4NL2), *B. subtilis* (3HSB), *S. aureus* (1KQ1), and *C. crescentus* (6GWK). However, the input to Rosetta FloppyTail is not just the crystal structure, but also requires the disordered regions to be in extended conformations ($\phi = -135°$, $\psi = 135°$). As IDRs are not typically resolved in crystal structures, I developed a PyRosetta[31] script (released with Rosetta in the public PyRosetta scripts directory under: `floppy_tail_utility/extend_terminus.py`) to append or prepend the missing residues. I used this script (*e.g.* `extend_terminus.py -c A -o 1hk9.chainA.pdb -p 1hk9.clean.pdb MAKGQ`) to add the N- and C-terminal residues for all of the above Hfq proteins (sequences can be found in Table 5.1). An additional PyRosetta utility script (`convert_to_beta.py`) to extend regions in a beta-strand conformation exists, if the residues are present in the input crystal structure and do not have to be added. Any mutants were generated using the PyMOL "mutate" function. Before modeling, the input structures with extended termini were "relaxed" with constraints using the FastRelax protocol[32,33], to eliminate energetically unfavorable atomic clashes.

Table 5.1: IDR sequences appended to Hfq crystal structures.

Species	NTD Sequence	CTD Sequence
E. coli	MAKGQ	SRPVSHHSNNAGGGTSSNYHHGSSAQNTSAQQDSEETE
P. aeruginosa	MSKGHS	SRPVRLPSGDQPAEPGNA
L. monocytogenes	MKQGGQG	SPQKNVALNPDAE
B. subtilis	MKPIN	PQKNVQLELE
S. aureus	MIANEN	VETEGQASTESEE
C. crescentus	MSAEKKQN	PAQPVQLYEPSADADD

5.3.2 FloppyTail modeling of IDRs

A modified version of the FloppyTail algorithm was used to model the disordered termini (see Appendix Figure 5.A.2). The FloppyTail algorithm generates hypothetical, low-energy

conformations of disordered regions through two stages of modeling: (1) low-resolution modeling, where side chains are represented as single pseudo-atom centroids, with aggressive sampling of backbone conformational space and gradient-based minimization, and (2) all-atom modeling, where all side-chain atoms are restored, with fine sampling of backbone conformational space, side-chain optimization, and minimization. At the end of each stage, the lowest-energy conformation is recovered. Non-disordered residues have no backbone motion, but are permitted to sample side-chain conformations.

I adapted the original algorithm to permit simultaneous modeling of multiple disordered termini. To this end, I expanded the underlying FoldTree. As detailed in Chapter 3, the FoldTree is a data structure within Rosetta that dfeines the order in which residue positions are updated. The default approach is to update from the N- to the C-terminus, mimicking protein folding. This is not optimal for modeling multiple disordered termini. Instead, I implemented a FoldTree that updates from the center-of-mass out towards the termini, for each polypeptide chain.

Another improvement I implemented was more extensive and better characterized sampling of the disordered energy landscape. To achieve this, I tracked the lowest energy observed at each step the low-resolution stage of an ultra-long *E. coli* Hfq[ii] simulation, assessing sampling in terms of energy and identifying an optimal number of low-resolution steps. The ultra-long simultaion reveal that the energy drops substantially ($\sim$360 Rosetta Energy Units [REU] for *E. coli* Hfq) in the first 100,000 steps of the low-resolution stage, but in the next 900,000 steps the change in energy is minimal ($\sim$60 REU). Following the ultra-long simulation, I repeated ten shorter low-resolution simulations to assess the stability of my observation, and the results showed that the ultra-long simulation is indeed representative (Figure 5.1). Based on this data and considering the computational cost, 50,000–100,000 steps or approximately 400 backbone moves attempted per disordered residue were identified as optimal, with the smaller step count used for the non-*E. coli* Hfq proteins with shorter IDRs.

[ii] I tested *E. coli* Hfq as it had the largest IDR, so the number of steps determined here should suffice for any other Hfq protein in my set.

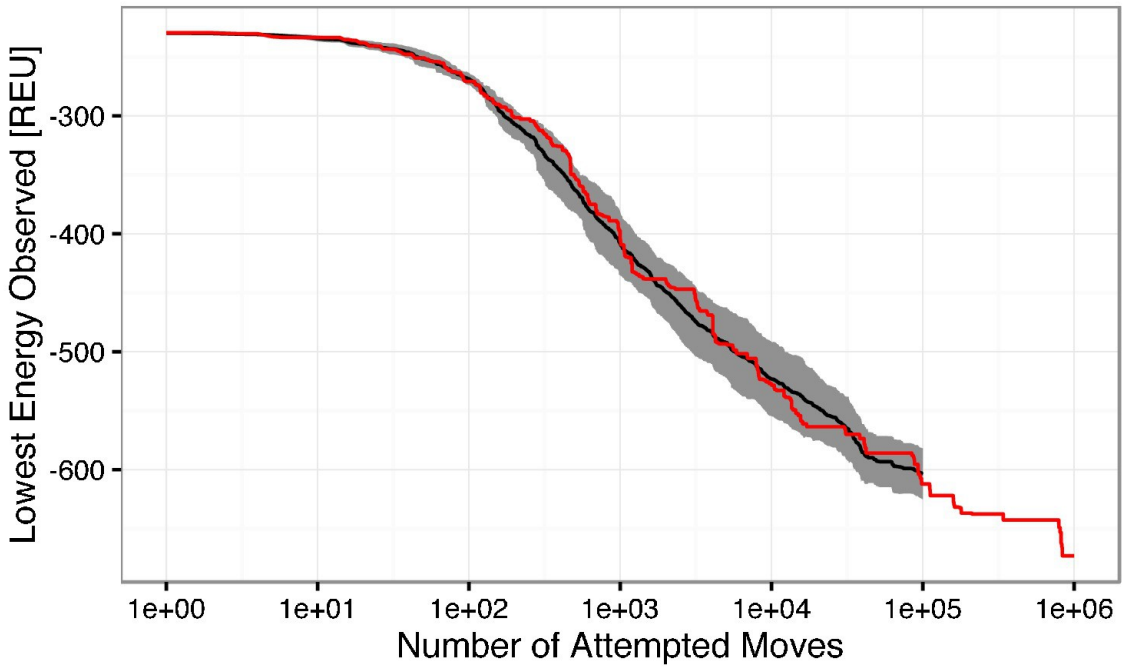

Figure 5.1: Semi-log plot of the lowest observed energy for a given number of attempted low-resolution backbone moves. The mean lowest energy observed for ten simulations is shown in black, with ± one standard deviation filled in gray. A single simulation run with 1,000,000 attempted moves is shown in red. The change in energy between 100,000 and 1,000,000 steps is ∼60 Rosetta Energy Units (REU), which is marginal in comparison to the ∼360 REU change in the first 100,000 attempted moves. The beneficial energy drop beyond 100,000 steps is outweighed by the computational cost. Thus, to generate one *E. coli* Hfq model, approximately 100,000 low-resolution moves should be sufficient to sample the energy landscape.

I repeated a similar analysis for the high-resolution stage, except there was no need for an ultra-long simulation as this stage converged within 100 steps (Figure 5.A.1). Ultimately, I elected to retain 1,000 high-resolution steps, as this was closer to the previously published 3,000 steps[12]. Based on the previous publication, simulations were used to generate a total 30,000 hypothetical structures for each species' Hfq protein, but only the 100 lowest-energy models were analyzed. The full FloppyTail algorithm is detailed in Figure 5.A.2.

5.3.3 Analysis of FloppyTail models

My final contribution to FloppyTail was an ensemble approach to model analysis. To this end, I used PyRosetta[31] to evaluate the energies of pairwise residue–residue interactions. Pairwise energies were computed with the `talaris2014` energy function[34], with terms

capturing van der Waals, solvation, hydrogen bonding and electrostatic interactions. If a pairwise energy was unfavorable (0 or greater), I did not consider it for further analysis. This script is publicly released with Rosetta in the PyRosetta scripts directory under: (`floppy_tail_utility/identify_interactions.py`.

To determine the nature of the CTD interactions, I considered two residue sets, those in the core (denoted $\mathcal{C}$, for all residues, and $\mathcal{B}$, if basic) and acidic residues in the tail ($\mathcal{T}$) (see Table 5.2 for the species-specific definitions). I calculated the average number of tail interactions for a single core residue, $x \in \mathcal{C}$, by counting the number of pairwise interactions, with a lower energy than a pre-determined threshold, between x and every residue in $\mathcal{T}$, then dividing by the total number of CTDs in the simulation:

$$\langle N_x \rangle = \sum_{\text{models}} \sum_{\text{subunits}} \sum_{y \in \mathcal{T}} \delta(x,y) / (N_{\text{models}} \cdot N_{\text{subunits}}),$$

where $\delta(x,y)$ is 1 if the residues are interacting ($E(x,y) < T$) and 0 if the residues are not interacting ($E(x,y) \geq T$), according to a threshold value, T, and the pairwise energiy $E(x,y)$.

Table 5.2: Core and tail residue selections for energy calculations.

Species	**Core ($\mathcal{C}$)**	**Basic Core ($\mathcal{B}$)**	**Acidic Tail ($\mathcal{T}$)**
E. coli	1–65	3, 16, 17, 19, 47	97, 99, 100, 102
P. aeruginosa	1–65	3, 5, 16, 17, 19, 47	94, 97
L. monocytogenes	1–65	2, 16, 17, 19, 35	100, 102
B. subtilis	1–65	2, 16, 17, 37	71, 73
S. aureus	1–65	10, 16, 41	65, 67, 99, 101, 102
C. crescentus	1–65	5, 6, 18, 19, 21, 49, 50	79, 80, 81

The threshold was determined by analysis of pairwise interactions. Figure 5.A.3 shows the energy distributions for all basic–acidic residue pairs across all *E. coli* Hfq simulations. There are two clear populations: one at 0 REU representative of non-interacting pairs and one at -2 REU representative of interacting pairs. Thus, the threshold is -1.0 REU for the `talaris2014` energy function. For the newest energy function, `REF2015`[35], the threshold is doubled to -2.0 REU the function weights have changed (data not shown).

The standard deviation for the average number of tail–core interactions with residue x was computed by using bootstrap resampling as previously described for protein docking[36]. Briefly, I resampled, with replacement, the set of models one thousand times ($B = 1,000$) and calculated resampled counts, N'_x, with the equation above for the resampled models. The standard deviation was computed as: $\sigma^2_N = \sum_B (N'_x - \langle N'_x \rangle)^2 / B$, where $\langle N'_x \rangle$ is the average count for the resampled set.

Similarly to how the counts were computed, I calculated the average energy for each interaction meeting the threshold criteria, summed over all residues in the tail set:

$$\langle E_{x:\mathcal{T}} \rangle = \sum_{\text{models}} \sum_{\text{subunits}} \sum_{y \in \mathcal{T}} \delta(x,y) E(x,y) / (N_{\text{models}} \cdot N_{\text{subunits}}).$$

The standard deviation for the interaction energy was computed without bootstrap resampling; the energy has a distribution within the set of models, whereas the presence of an interaction is binary and only varies when the models are resampled. I computed the standard deviation as:

$$\sigma^2_E = \sum_{\text{models}} \sum_{\text{subunits}} \left(\sum_{y \in \mathcal{T}} \delta(x,y) E(x,y) - E_{x:\mathcal{T}} \right)^2 / (N_{\text{models}} \cdot N_{\text{subunits}}).$$

Finally, I defined an expected energetic contribution (EEC) metric, as the average energy of a tail–core interaction multiplied by the average number of that tail–core interaction per model: $\sum_{\mathcal{B}} \langle N_x \rangle \langle E_{x:\mathcal{T}} \rangle$. The standard deviation for EEC was computed by assuming that the standard deviations of the interaction counts and energies are independent: $\sigma^2_{\text{EEC}} = \sigma^2_E \sigma^2_N + \sigma^2_E \langle N_x \rangle^2 + \sigma^2_N \langle E_{x:\mathcal{T}} \rangle^2$.

5.3.4 Hfq purification and CTD binding studies

This work was done by Dr. Andrew Santiago-Frangos. Untagged *E. coli* Hfq102, Hfq-sCTD, Hfq65, Hfq65-Q35A, Hfq65-K47A, Hfq65-R19D and Hfq65-R16A were over-expressed in *E. coli* BL21(DE3)Δhfq::cat-sacB cells grown in 1 L LB-Miller media (10 g/L Tryptone,

10 g/L NaCl, 5 g/L yeast extract) supplemented with 100 μg/mL ampicillin. Plasmids for over-expression of mutant Hfq proteins were created by site-directed mutagenesis of pET21b-Hfq[15]. The purification method has been previously described[30]. In brief, resuspended cell lysates of Hfq102 and Hfq-sCTD variants were clarified by heat denaturation and untagged Hfq was purified via Ni^{2+}-affinity. Lysates of Hfq65 variants were further clarified by ammonium sulfate precipitation after heat treatment, and the protein purified by hydrophobic interaction chromatography. Finally, all Hfq variants were purified by cation-exchange chromatography to remove nucleic acids.

To measure binding of CTD-FITC, CTDpos-FITC, or BsCTD-FITC peptides (Table 5.3) to Hfq65 or Hfq65 mutants, the fluorescence polarization of FITC-labeled peptide was measured 3 min after the addition of 0–0.3 μM Hfq65. Anisotropy measurements were normalized to the average anisotropy in the absence of Hfq. Samples were prepared in a 100 μL cuvette containing 100 μL 50 mM TrisHCl pH 7.5, 45 nM CTD-FITC or CTDpos-FITC, at 30°C. Fluorescence polarization with grating correction factor was measured using a Horiba Fluorolog-3 (L-format) with single excitation and emission monochromators at 495 nm and 515 nm respectively (5 nm slit widths). Titrations were performed in duplicate and the curves were fit to a single-site binding isotherm: $y = K_a \cdot x/(1 + K_a \cdot x)$, in which K_a is the association constant.

Table 5.3: Sequences of peptides and RNAs.

Target	GUGGUCAGUCGAGUGG
Target-A18	GUGGUCAGUCGAGUGGAAAAAAAAAAAAAAAAA
Molecular beacon	FAM-GGUCCCCCACUCGACUCACCACCGGACC-DABCYL
CTD-FITC	FITC-Ahx-NNAGGGTSSNYHHGSSAQNTSAQQDSEETE-COOH
CTDpos-FITC	FITC-Ahx-NNAGGGTSSNYHHGSSAQNTSAQQRSNKTN-COOH
BsCTD-FITC	FITC-Ahx-QLELE-COOH

5.3.5 RNA binding and annealing

This work was done by Dr. Andrew Santiago-Frangos. The sequences RNA substrates are listed in Table 5.3. The purification protocols for the molecular beacon[25] and Target/Target-

A18[37] have been previously described. Annealing kinetics of molecular beacon (50 nM) to either Target or Target-A18 RNA (100 nM) by 0–200 nM Hfq hexamer, in 1X TNK (10 mM TrisHCl pH 7.5, 50 mM NaCl, 50 mM KCl) buffer at 30 °C, were measured by stopped-flow fluorescence spectroscopy as described previously[19,25]. Annealing progress curves were fit to single or double-exponential rate equations.

5.3.6 Hfq alignments and sequence logos

This work was done by Dr. Andrew Santiago-Frangos. All Hfq gene sequences were taken from Uniprot[38]. 5359 sequences were aligned using the G-INS-I algorithm on MAFFT webservers[39]. This alignment was reduced using CD-HIT[40] and Max-Align[41]. An unrooted, neighbor-joining tree of the remaining 985 non-redundant, representative, sequences was made on MAFFT webservers[39]. Sequence logos of re-aligned sequences from chosen clusters were generated using WebLogo[42].

5.4 Results

5.4.1 C-terminus of Hfq is enriched for acidic residues

To search for conserved features or amino acid motifs amongst the highly heterogeneous Hfq CTDs, a phylogenetic tree was constructed from the multiple sequence alignment of nearly 1000 non-redundant sequences. The cluster containing *E. coli* Hfq contained many other Hfq variants previously identified as functional in RNA annealing[26] or sRNA regulation[43]. Therefore, the sequence logo of this cluster of 222 Hfqs was examined in more detail Figure 5.2A.

The start of the CTD region is delineated by a proline at position 64 of *E. coli* Hfq that is strongly conserved across all clades. Additionally, an arginine at the beginning of the CTD (position 66 in *E. coli*) that packs against the lateral edge of the Hfq hexamer[44] is strongly conserved. Although the middle linker region of the CTD lacks conserved motifs, the C-terminus is rich in acidic residues, corresponding to the sequence DSEETE in *E. coli*.

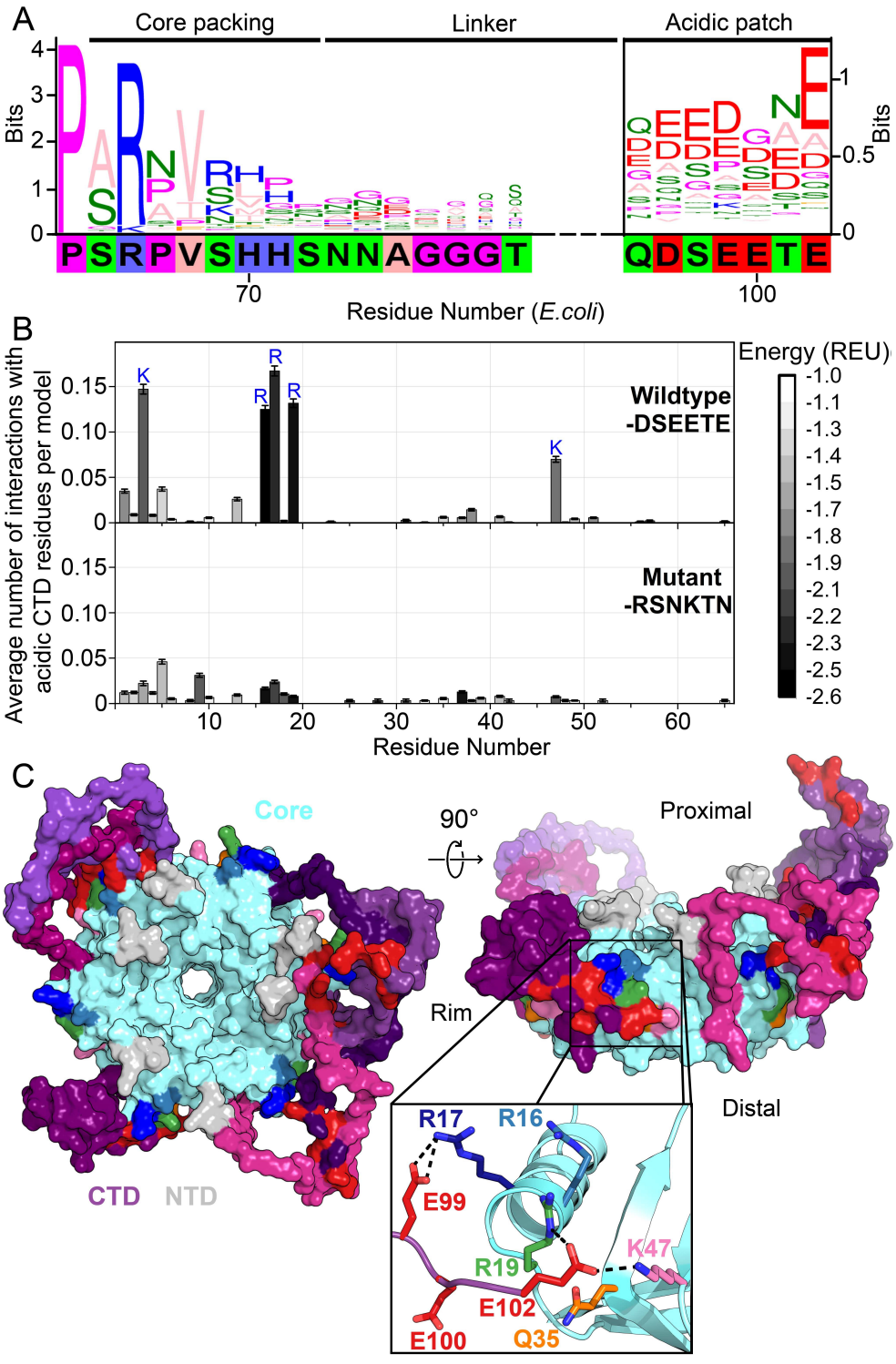

Figure 5.2: Caption follows on the next page.

Figure 5.2: (Previous page.) (A) Sequence logo of the CTD generated from gapped alignment of Hfq sequences that clustered with *E. coli* Hfq, numbered according to the E. coli sequence. Regions of interest are denoted above. The gapped *E. coli* CTD sequence is shown below. (B) (Top) Average number of times a given core residue favorably interacts ($E < 1.0$ REU) with at least one acidic CTD residue, per low energy model. Acidic CTD residues most frequently interact with basic Hfq core residues. (Bottom) Mutation of acidic CTD residues 97, 99, 100 and 102 to basic or polar residues decreases the number of predicted core interactions. Error bars represent $\pm$ 1 s.d. as computed by bootstrap resampling of the computational models. Of 36 core residues not predicted to interact with the CTD, 14 had accessible surface area < 2.0 A^2, computed in PyMOL. (C) (Left) Example low-energy model of wildtype *E. coli* Hfq; top-down proximal view. Light grey, NTD; cyan, Hfq core; pink-purple, CTD; red, CTD tip. (Center) Side view of rim of the same Hfq model. (Inset) Example hydrogen bonding network at the CTD–core binding interface showing interactions between the acidic CTD residues (red) and core residues as indicated.

Noting that most Hfq clusters containing a basic patch on the rim also end in acidic residues, it was hypothesized that the CTD tip binds the rim. Because the basic patch is essential for sRNA binding and annealing, direct interaction between the CTD tip and the Hfq core could explain the previously observed auto-inhibition of the CTD[30].

5.4.2 De novo modeling of CTD interactions in the Hfq hexamer

To determine whether the acidic tip of the E. coli Hfq CTD could interact with basic residues in the core, I used Rosetta FloppyTail[12], a *de novo* modeling approach for disordered regions of proteins. I updated the original FloppyTail algorithm to model multiple disordered regions simultaneously and to ensure adequate sampling of backbone degrees of freedom. Then, I generated and analyzed 30,000 models of the full-length *E. coli* Hfq hexamer. In the lowest energy (1%) subset of models, the acidic CTD residues (D97, E99, E100, and E102) frequently interact with basic residues on the rim (R16, R17, R19 and K47) and in the NTD (K3) (Figure 5.2B, top). By contrast, K31 on the distal face is not predicted to be contacted by the CTD, although K31 is highly accessible. This bias accords with prior observations that the CTD does not displace RNA from Hfq's distal face[30]. As anticipated for a disordered domain, no single conformation dominated the ensemble of models (Figure 5.A.4). Rather, the acidic CTD tip was found to bind to various combinations of residues in the basic patch (Figure 5.2C).

To confirm that I was not simply observing the non-specific collapse of the disordered CTD onto the core, or enriching interactions between highly solvent-accessible polar residues, I repeated these simulations using a mutant Hfq in which the acidic CTD residues were replaced with polar or basic side chains (D97R-E99N-E100K-E102N). These mutations drastically decreased the frequency of predicted interactions between the basic core residues and CTD residues at positions 97, 99, 100 and 102 in our simulations (Figure 5.2B, bottom), without increasing predicted interactions between this mutant CTD and solvent-accessible acidic residues on the Hfq core (D9, E18, E37 and D40).

5.4.3 Acidic CTD specifically binds Hfq rim

To determine whether the CTD interacts with the rim as predicted by our models, we used fluorescence anisotropy to measure the affinity of core Hfq (Hfq65) for a fluorescently-labeled CTD peptide, CTD-FITC (Figure 5.3A). CTD-FITC lacks residues 65–72 to avoid contributions to binding from this region, which packs against the Sm domain as one strand of the β-sheet. Hfq65 bound to CTD-FITC with a K_d of 2.9 μM Hfq monomer in low salt buffer (cyan in Figure 5.3B). Binding of the CTD-FITC peptide to Hfq65 was weakened by mutations in the basic rim residues R16A, R19D and K47A (Figure 5.3B), which frequently interact with the CTD in the computational models (Figure 5.2B). In contrast, mutation of a surface-accessible polar residue (Q35A) close to the binding interface (Figure 5.2C, inset), slightly enhanced CTD binding (Figure 5.3B). Intriguingly, A35 is common in Hfq from γ-proteobacteria. Finally, a CTD peptide containing the mutated C-terminal tip (RSNKTN) was not able to bind Hfq65, confirming that the acidic residues on the CTD peptide are necessary for this interaction (grey in Figure 5.3B).

5.4.4 Low-scoring FloppyTail models identify key CTD interactions

To determine how much core residues that bind the CTD contribute to Hfq's RNA annealing activity, we compared the effect of rim mutations on the rate of base pairing between an

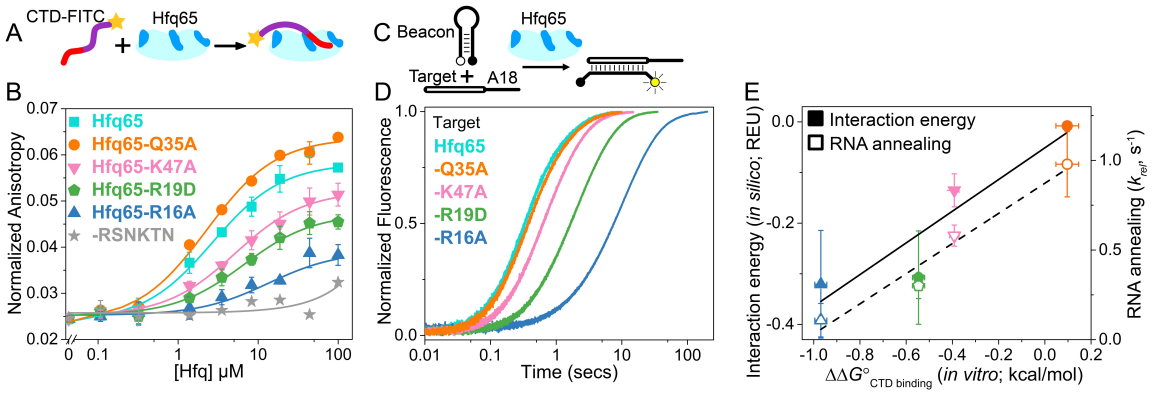

Figure 5.3: (A) Scheme for in vitro binding of fluorescent CTD-FITC peptide by Hfq core. The CTD linker is shown in purple, the acidic tip in red and the N-terminal FITC as a yellow star. Hfq core is shown in cyan, with basic rim patches in dark blue. (B) Binding of CTD-FITC to variants of Hfq65 core at 30 °C. 45 nM CTD-FITC was titrated with 0–100 μM Hfq monomer in duplicate, and the average ($\pm$s.d.) was fit to a single-site binding isotherm. (C) Reaction scheme for annealing an RNA molecular beacon to a target RNA (open bar)[25]. (D) Progress curves for annealing 50 nM molecular beacon and 100 nM Target by 50 nM Hfq65 hexamer at 30 °C, measured by stopped-flow fluorescence. (E) Contribution of core residues to CTD binding. Interaction energy (Expected Energetic Contribution; EEC) *in silico* for a core residue in the Rosetta models (solid symbols and solid line; adjusted $R^2 = 0.77$) or the average annealing rates for Target and Target-A18 relative to Hfq65 (open symbols and dashed line; adjusted $R^2 = 0.94$) versus experimental CTD binding energy ($\Delta\Delta G^\circ$) for each Hfq65 variant. The binding energy, $\Delta\Delta G^\circ = -RT \ln(K_d^{\mathrm{MUT}}/K_d$, reflects the perturbation to CTD binding by a mutation in Hfq65. The interaction energy *in silico* or EEC is defined as the average energy of a tail–core interaction multiplied by the average number of tail–core interactions per model (Figure 1B, top and Equation 3). The relative annealing rate for Hfq65 variants, $k_{\mathrm{rel}} = k_{\mathrm{obs}}^{\mathrm{MUT}}/k_{\mathrm{obs}}^{\mathrm{WT}}$, is <1 if the mutated residue is important for RNA annealing.

RNA molecular beacon and a 16 nt Target RNA by stopped-flow fluorescence spectroscopy (Figure 5.3C). In the absence of competition from the CTD, the rate of annealing in this assay depends only on interactions between the two RNAs and the Hfq core. As previously observed[30], Hfq65 is highly active in single-turnover annealing assays (Figure 5.A.5A). The observed annealing rate was most diminished by the loss of basic residues, especially the conserved R16A, and relatively unaffected by the mutation Q35A (Figure 5.3D). Similar results were obtained with Target-A18, which anchors to the distal face (Figure 5.A.5B). The average relative annealing rates of Hfq65 variants correlated well with the importance of each residue for CTD binding in vitro (Figure 5.3E), suggesting that the CTD peptide and the RNA interact with the same residues on the rim of Hfq.

The predictive value of our computational approach was validated by a direct correlation between the experimentally measured contribution ($\Delta\Delta G^{\circ}$) of each core residue for CTD binding with the predicted Expected Energetic Contribution (EEC) of that core residue to interactions with the acidic CTD in silico (solid symbols and solid line, Figure 5.3E). EEC is defined as the average energy of a tail–core interaction multiplied by the average number of tail–core interactions per model. The absolute binding and simulated interaction energies cannot be directly compared because the peptide binding assay is performed in trans rather than in cis, and the Rosetta Energy does not account for entropic contributions to binding. Nevertheless, amino acids that most strongly impacted the free energy of CTD binding when mutated, also had larger contributions to CTD binding in silico (solid symbols and solid line, Figure 5.3E; linear regression p-value=0.078), and had stronger effects on Hfq65 RNA annealing activity in vitro (open symbols and dashed line, Figure 5.3E; linear regression p-value=0.020).

5.4.5 Key CTD interactions correlate with activity in other species

The results for *E. coli* Hfq show that the strength and frequency of CTD–core interactions depend on the number of basic residues in the core, the acidic residues in the CTD, and

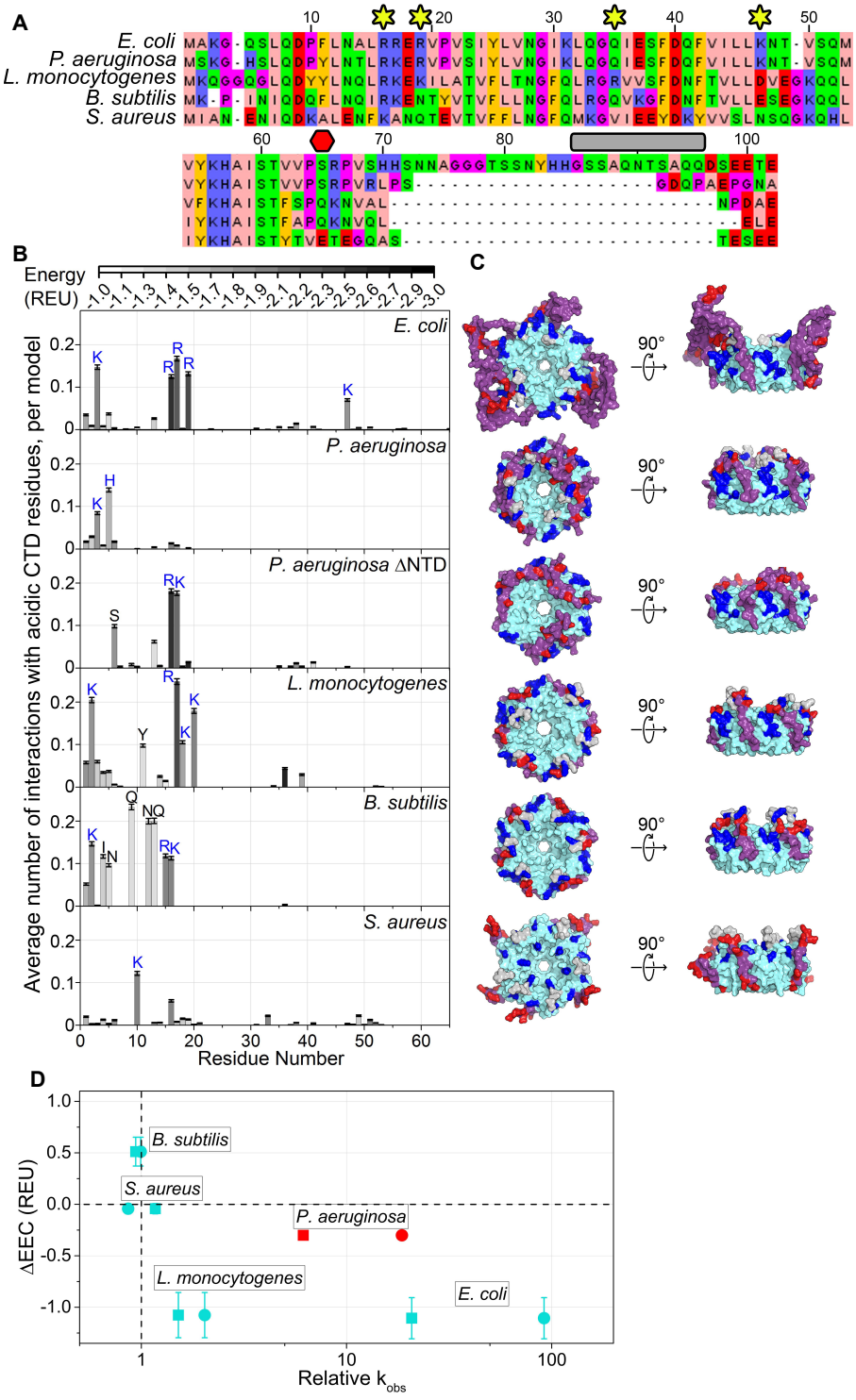

Figure 5.4: Caption follows on the next page.

Figure 5.4: (Previous page.) (A) Alignment of modeled Hfq sequences in order of decreasing *in vitro* RNA annealing activity. Residues are numbered according to the *E. coli* sequence. Yellow stars, residues mutated in this study; red hexagon, last residue in Hfq65; grey box, linker removed in Hfq-sCTD. (B) Average number of favorable interactions per model for each core residue with at least one acidic CTD residue in the lowest energy models ($\leq$1%). As in Figure 5.2B. Number of residues with $< 2.0A^2$ accessible surface area: *P. aeruginosa*, 10; *L. monocytogenes*, 25; *B. subtilis*, 19; *S. aureus*, 12. (C) Top-down (proximal face) and side (rim) views of example low-energy models for each Hfq, as in Figure 5.2C. (D) Relative RNA beacon annealing rate for Target-U6 (boxes) and Target-A18 (circles) in Hfq vs. no Hfq (relative k_{obs}) versus the specificity of predicted CTDâĂŞcore interactions (ΔEEC) for *B. subtilis*, *S. aureus*, *L. monocytogenes* and *E. coli* Hfq (blue), and *P. aeruginosa* Hfq, which is more active in vitro than predicted by its ΔEEC (red). Annealing data are from Zheng *et al.*[26].

the linker length. Thus, the proposed mechanism for CTD–core interactions can be used to predict how the degree of CTD autoinhibition may vary among bacterial Hfq's. I applied my *de novo* modeling procedure to estimate the CTD–core interactions in four other bacterial Hfqs (Figure 5.4A) for which the genetic function and *in vitro* annealing activity have been previously characterized[26,45–49]. I examined low energy models of each Hfq hexamer, and compared how frequently acidic CTD residues interact with basic rim and NTD residues ("on-target") versus other core residues ("off-target") (Figure 5.4B,C). This comparison was quantitatively expressed as the difference in the EEC of on-target and off-target interactions (ΔEEC).

For *E. coli* Hfq, an active chaperone with a basic rim patch and long CTD, the CTD tip tended to interact with basic residues on the rim and NTD more often and more strongly than with other residues, resulting in $\Delta EEC = 1.11 \pm 0.20$ REU. This was also true for *Listeria monocytogenes* Hfq ($\Delta EEC = 1.08 \pm 0.22$ REU). In contrast, Bacillus subtilis ($\Delta EEC = 0.51 \pm 0.14$ REU) and *Staphylococcus aureus* ($\Delta EEC = 0.19 \pm 0.05$ REU) Hfq, which are inactive in our *in vitro* annealing assay[26], did not exhibit specific CTD–core interactions. Finally, in models of full-length *Pseudomonas aeruginosa* Hfq, the CTD adopts an extended β conformation that wraps over the rim of the hexamer and places the C-terminal acidic residues near the weakly basic NTDs ($\Delta EEC = 0.30 \pm 0.04$ REU) (Figure 5.4C). In the absence of the NTD, however, the CTDs dock with R16 and K17 on the rim ($\Delta EEC = 0.50 \pm 0.18$ REU). Thus, Hfqs that do not anneal RNA *in vitro* tend to possess

shorter, less acidic CTDs that form weaker and less frequent interactions with the basic rim and NTD in silico (Figure 5.4D). There is a similar trend between ΔEEC and the importance of Hfq for sRNA regulation in each bacterium[14,45–47,49–52].

In the above examples, both the CTD and the core co-vary between different species. I next asked whether the CTD conferred specificity or strength to CTD–core interactions. I modeled an Hfq chimera consisting of the highly basic *E. coli* Sm core, fused to the shorter and slightly less acidic *B. subtilis* CTD. In our models, the *B. subtilis* CTD contacted K3 in the NTD and R17 on the rim more frequently than *E. coli* CTD (Figure 5.5A), but contacted R16 and R19, which are functionally very important (Figure 5.2 and Figure 5.3), less frequently than *E. coli* CTD (Figure 5.5A). This was corroborated with fluorescence anisotropy results showing that *E. coli* Hfq65 binds a BsCTD-FITC peptide about three times more weakly than its own CTD (8.7 μM vs. 2.9 μM; Figure 65.5B). Although it was shown that a foreign CTD can bind the core of *E. coli* Hfq, the "specificity" of this interaction may have been lost.

5.4.6 FloppyTail ensembles capture crystallizable states

So far, FloppyTail has demonstrated an ability to model interaction energies in the *E. coli* Hfq CTD with good correlations to *in vitro* mutagenesis assays, and to be potentially predictive of *in vitro* annealing activity of four other Hfqs. So, I next sought structural validation, in collaboration with Dr. Steven Hardwick (Department of Biochemistry, Cambridge University). Dr. Hardwick had recently acquired a crystal structure of full-length *C. crescentus* Hfq, with sufficient electron density to model the CTD (PDB ID 6GWK).

To ensure that the results were not biased, the calculations were performed without prior knowledge of the crystal structure. The overall basicity of the rim is conserved between *C. crescentus* and *E. coli* Hfqs 5.2. However, the arginines are distributed more towards the rim-distal face in *C. crescentus* Hfq. Among the lowest-energy fraction of models generated in the simulations (1% of all models, sorted by energy), acidic residues at the CTD of *C. crescentus* Hfq were found to frequently form energetically favourable contacts with basic

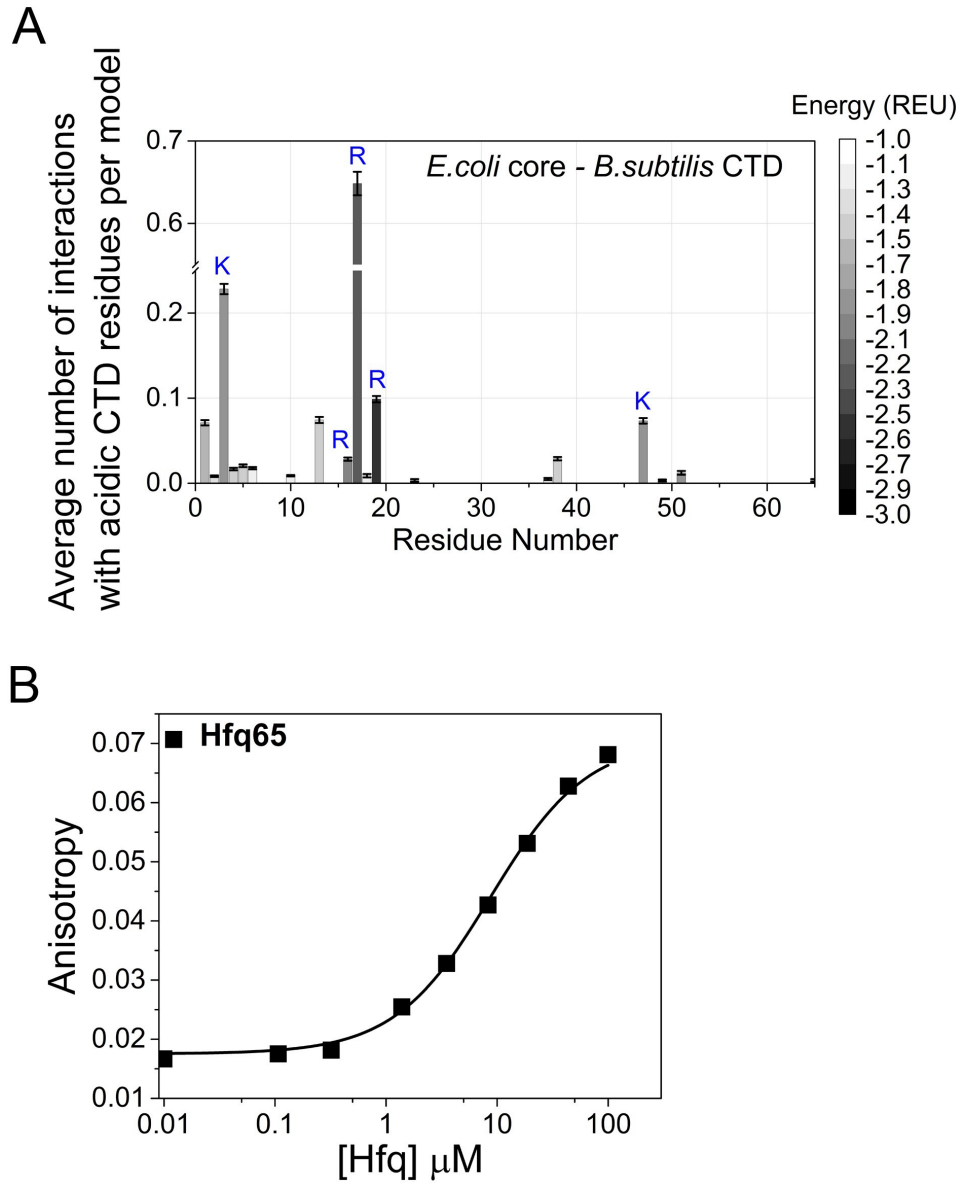

Figure 5.5: A chimera of *E. coli* Hfq core (residues 1–64 *E. coli* numbering) and the *B. subtilis* CTD (residues 65–73 E. coli numbering; QKNVQLELE) was modeled with Rosetta FloppyTail as in Figure 5.2. (A) Average number of energetically favorable interactions ($E < 1.0$ Rosetta Energy Units) with at least one acidic CTD residue, per low energy model as in Figure 5.2. Error bars ($\pm$1 s.d.) computed by bootstrap resampling. 14 of the core residues not predicted to interact with the CTD were solvent inaccessible. (B) Binding of BsCTD-FITC peptide to *E. coli* Hfq65 core at 30 °C. 45 nM BsCTD-FITC was titrated with 0–100 μM Hfq monomer in duplicate, and the average ($\pm$ s.d.) was fit to a single-site binding isotherm with $K_d = 8.7\mu$M. Although the binding strength is three times weaker than for the *E. coli* CTD peptide, the rank order for interactions with basic core residues no longer coincides with their relative importance for RNA annealing *in vitro* (Figure 5.3E). Therefore, the *B. subtilis* CTD may not be optimal for auto-regulating the core of *E. coli* Hfq.

residues arginine 18, lysine 19 and lysine 21 on the proximal-rim interface (Figure 5.6A), even when the basic NTDs were excluded from the model. By contrast, few contacts were observed to arginines 49 and 50, which are solvent exposed but lie toward the distal side of the rim. Nearly all the lowest-energy fraction of models had at least one CTD in contact with the rim, and many of the modelled CTD conformations closely resembled ($\sim$2 Å Cα RMSD) the crystallized CTD-rim interaction (Figure 5.6B and 5.A.6). *In silico* mutation of both arginine 18 and lysine 19 to alanine ablated interactions of acidic CTD residues with these positions on the core (Figure 5.A.6). These results illustrate the predictive power of Rosetta FloppyTail and suggest that the crystallized CTD–rim interaction is likely to occur in solution rather than being an artifact of crystal lattice packing.

5.5 Discussion

It was previously found that the flexible CTD of *E. coli* Hfq sweeps RNAs from the proximal and rim surfaces of the Hfq ring by an unknown mechanism[30]. Because the mechanism was not known, it was not possible to predict whether other bacterial Hfq CTDs, which are highly variable in sequence composition and length, would perform similar functions. In this chapter, computational models and experiments showed that the acidic tip of the CTD interacts with basic patches on the rim of Hfq, potentially competing with RNA *in vivo*. The good agreement between the modeled CTD–core contacts and the contributions of individual residues to the measured CTD binding energies and to RNA annealing validated the modeling approach, and further suggested that nucleic acids and the acidic tip of the CTD interact with the same residues in the Hfq core. As expected for a nucleic acid mimic, CTD–core interactions were dominated by electrostatics (Figure 5.3).

The Rosetta FloppyTail algorithm, enabled by my advances, provided atomic-scale insight to the accessible conformations of the disordered N- and C- termini of *E. coli* Hfq (Figure 5.2C). Furthermore, I developed a metric, EEC, for assessing disordered–order region interactions that correlated well with experimental measurements (Figure 5.3), with

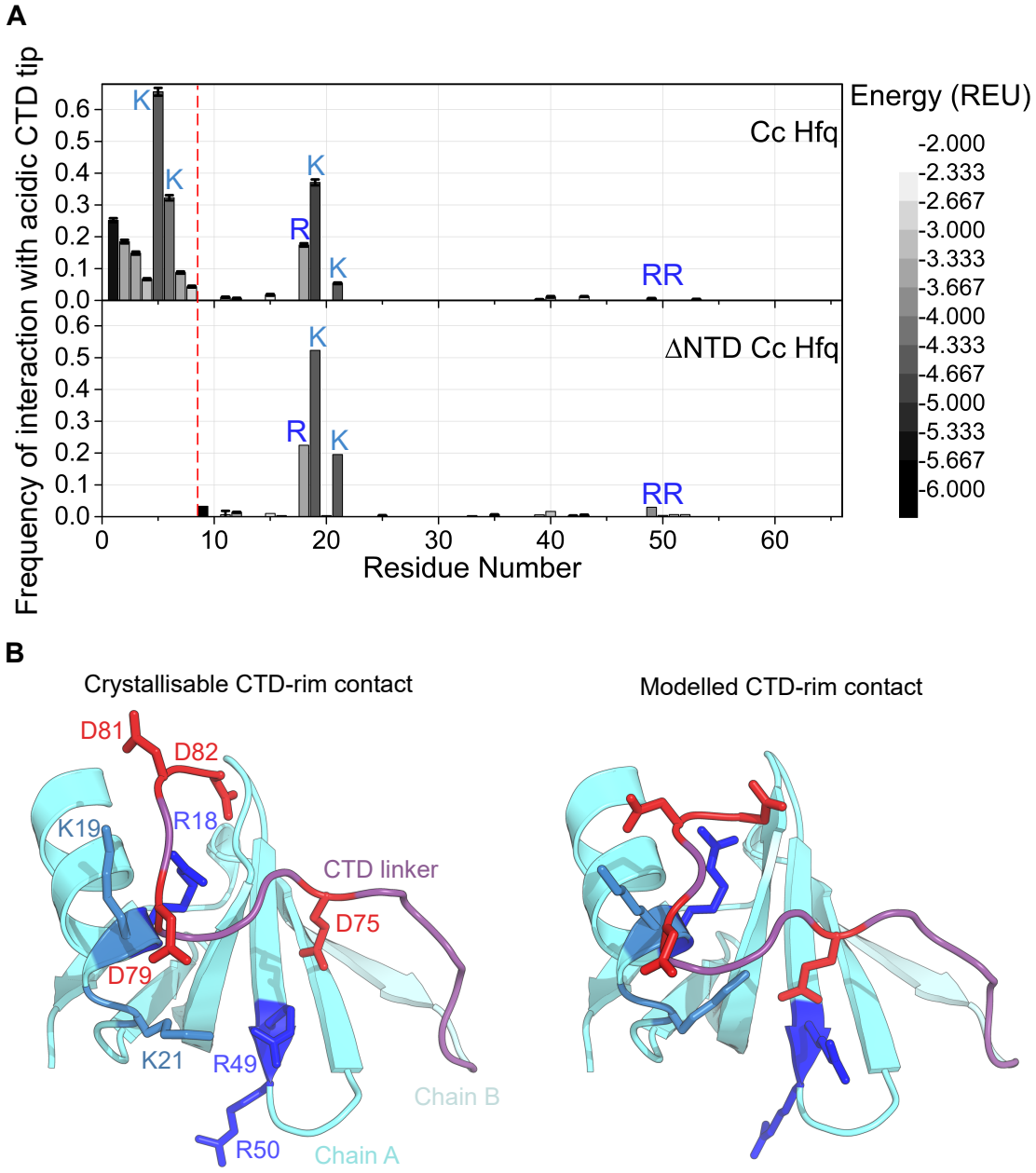

Figure 5.6: (A) Observed frequency of favourable ($E < -2.0$ REU) residue-to-acidic-CTD interactions in low-energy FloppyTail models. Top panel; full length Cc Hfq. Lower panel; Cc Hfq with residues of the Nterminal extension deleted (ÎŤNTE). Upon excluding the NTD *in silico*, only the basic core residues retain interactions with the CTD. Error bars show $\pm$ one standard deviation, as computed by bootstrap resampling. (B) Side-byside comparison of crystallized and modelled CTD-core contacts. The acidic tip residues in the CTD of the adjacent monomer (Chain B, light cyan) contact the basic rim residues of the monomer in the forefront (Chain A, cyan). For clarity, the NTE and CTD of chain A, and most of chain B except the last betasheet and CTD are omitted. CTDs, purple; lysines, skyblue; arginines, royal blue. CTD acidic residues are colored in red..

more accuracy than commonly used distance cutoffs (*e.g.* defining residues as interacting if Cα–Cα or Cβ–Cβ $< d$)[12]. The ease of modeling Hfq CTDs with FloppyTail enabled the study of different bacterial Hfqs. This *de novo* modeling strategy was able to identify frequent and specific CTD–rim interactions in *L. monocytogenes* and *P. aeruginosa* Hfq, which act in sRNA regulation and annealing, but not for *B. subtilis* and *S. aureus* Hfq (Figure 5.4D), in agreement with *in vitro* experiments. Finally, FloppyTail was found to recapitulate interactions observed *in crystallo* for *C. crescentus* Hfq. Together, these results suggest that the FloppyTail algorithm could be generally useful for predicting the interactions of disordered regions with ordered domains.

Many RNA and DNA binding proteins contain disordered or flexible domains that have been implicated in cooperativity, autoinhibition and liquid phase separation[53–55]. Hfq is an example of an emerging paradigm of autoregulation of nucleic acid binding by nucleic acid mimic peptides. Other examples in which a disordered CTD autoinhibits RNA or DNA binding include HTLV-1 NC[56], *E. coli* gyrase[57], *E. coli* ssDNA binding protein[58] and mammalian high-mobility group B1[59]. Unlike HTLV-1 NC, which also remodels RNA, the Hfq CTD gives rise to dynamic cycling of bound RNAs needed to chaperone sRNA–mRNA interactions. Our modeling procedure could be utilized to screen disordered domains found in kinases, such as myosin light chain kinases and protein kinase C[60], and nucleic acid binding proteins from all kingdoms of life.

5.A Appendix

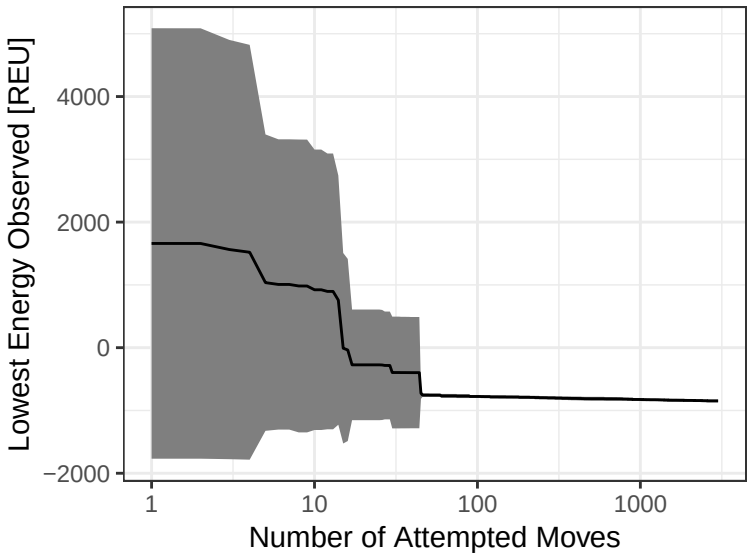

Figure 5.A.1: Semi-log plot of the lowest observed energy for a given number of attempted high-resolution moves. The mean lowest energy observed for five simulations is shown in black, with $\pm$ one standard deviation filled in gray. The initial broad range of high energies is due to clashes introduced by replacing centroid representations of side chains with all-atom ones. Within 100 moves, the simulations have converged, therefore only 100 high-resolution moves are attempted in each simulation.

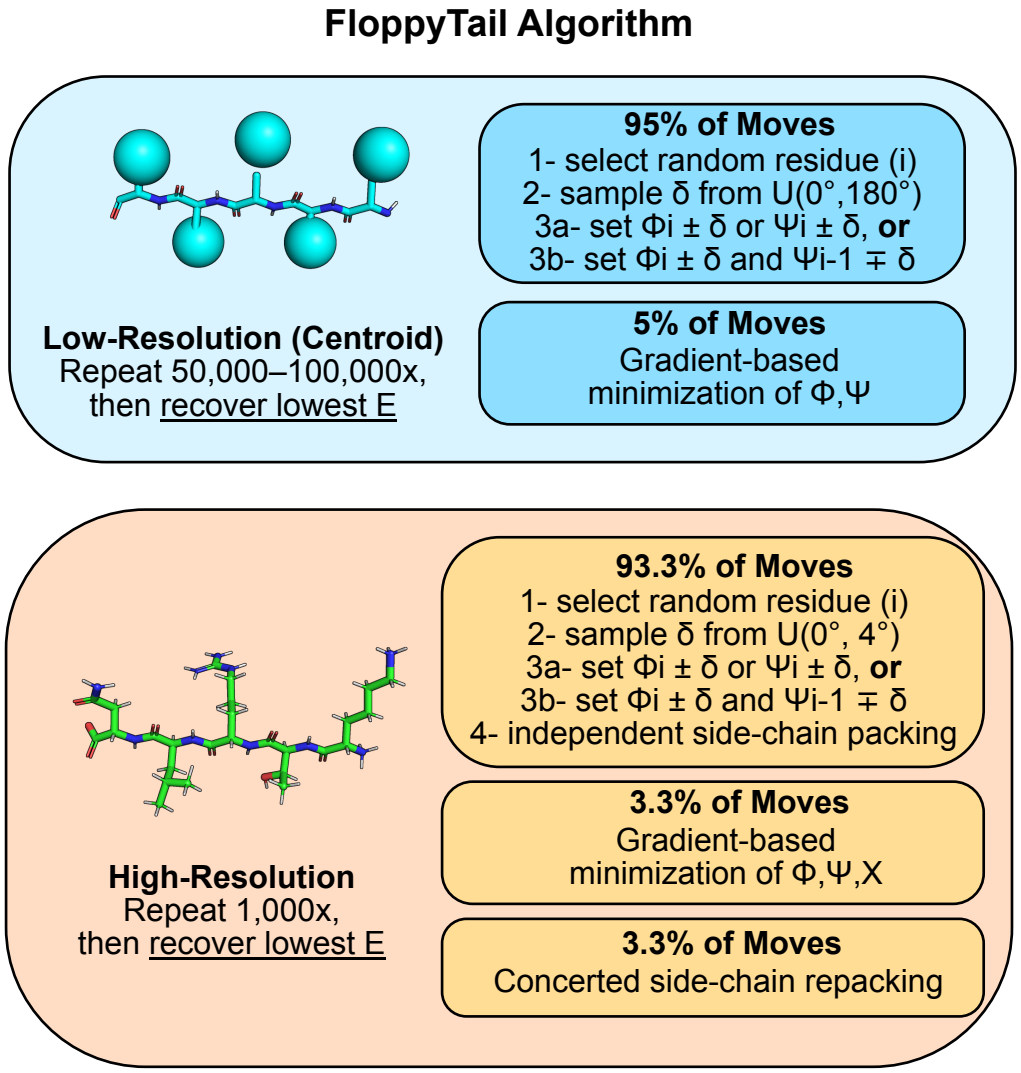

Figure 5.A.2: The FloppyTail algorithm operates in two stages: low-resolution (centroid) and high-resolution. In the centroid stage, 50,000–100,000 moves are attempted. 95% of the moves are either (3a, 47.5%) Small or (3b, 47.5%) Shear moves whereas the remaining 5% are minimization moves. A Small move randomly perturbs either ϕ or ϕ by up to 180°. A Shear move randomly perturbs ϕ or ϕ by up to 180°and then makes an equal and opposite compensatory move in the preceding/following angle. The lowest-energy conformation is used as input for the high-resolution state. In the high-resolution stage, all side-chain atoms are restored and only 1,000 moves are attempted. 93.3% of the moves are Small or Shear, but with a smaller magnitude (4°) followed by RotamerTrials, or independent sampling of χ angles for each side chain, 3.3% of moves are minimization moves, and the final 3.3% of moves are PackRotamers moves, or concerted side-chain repacking (where *chi* angles of multiple positions are simultaneously optimized). At the end of the simulation, the lowest-energy conformation is ouptut. In a typical simulation, 30,000 models will be generated.

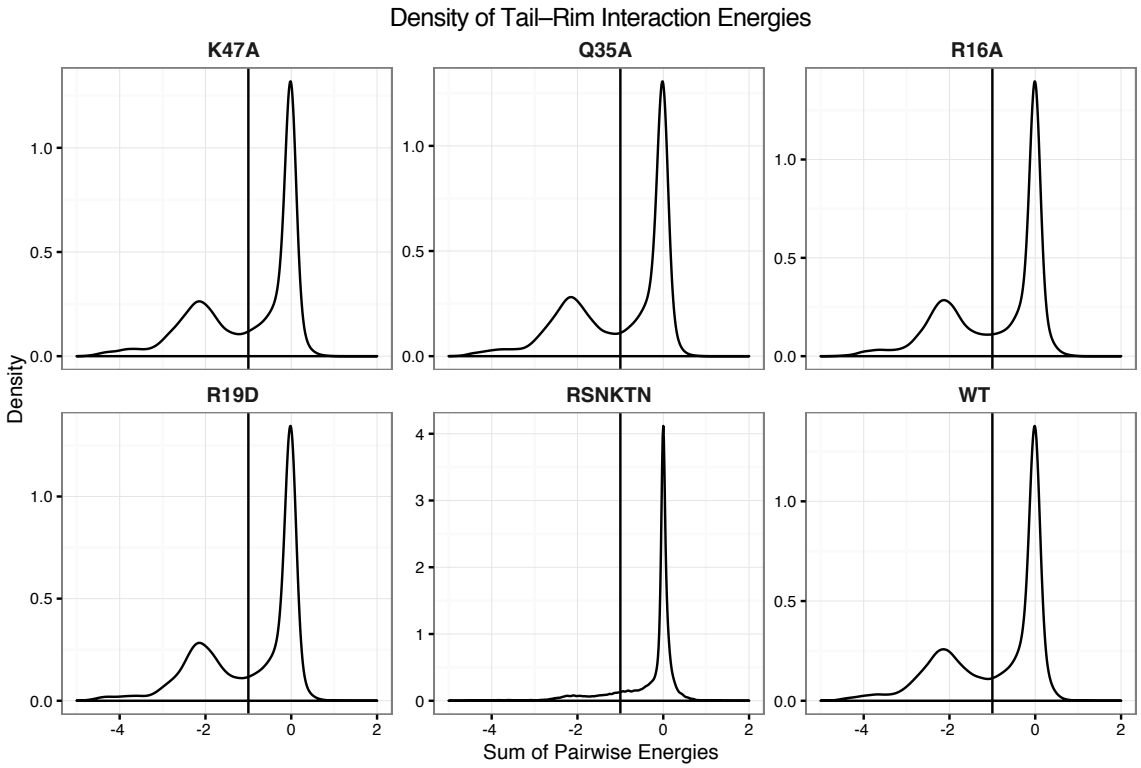

Figure 5.A.3: Energy densities of the pairwise negative-tail to positive-core residue–residue interactions in all models. The vertical line represents the cutoff used to define an "interacting" pair. Each distribution has two "bumps": one at 0 REU for non-interacting residues and one at -2 REU for interacting residues.

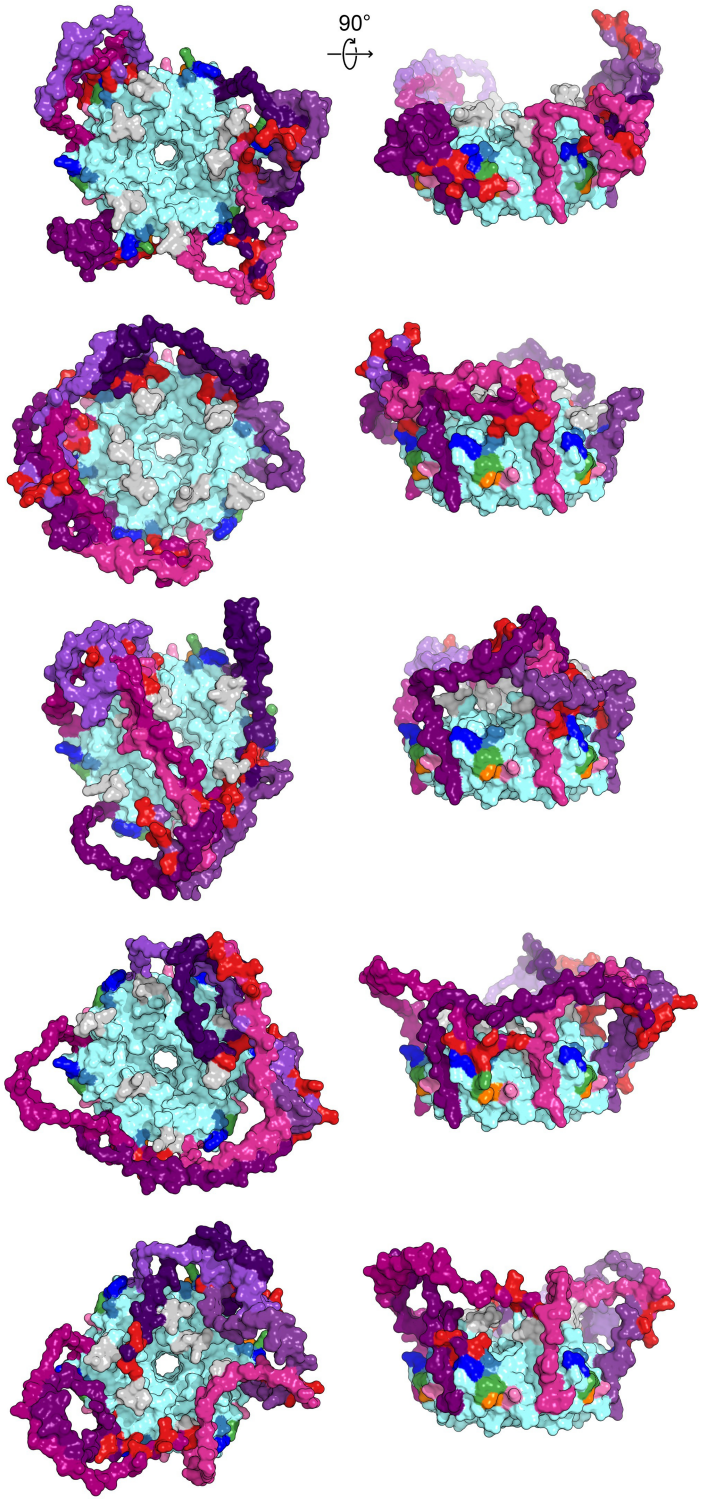

Figure 5.A.4: A gallery of additional low-energy E. coli Hfq models from either a top-down view through the proximal pore (right) or side-on view of the rim (left), colored as in Figure 5.2.

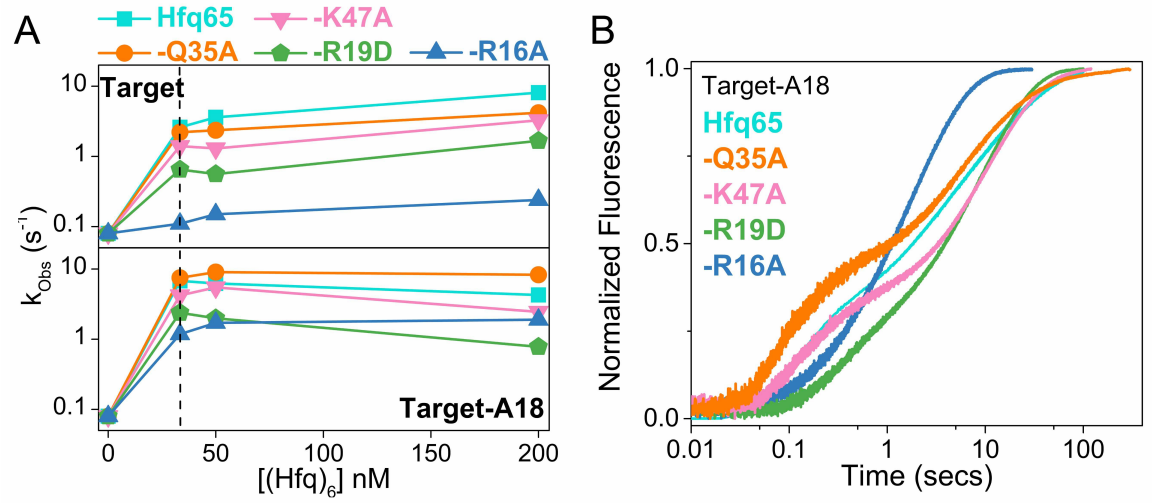

Figure 5.A.5: Annealing of 100 nM Target RNA to 50 nM beacon RNA by Hfq65 and Hfq65 variants was measured by stopped flow fluorescence at 30 °C in 1 ÃŮ TNK buffer. (A) Observed annealing constants with 0–200 nM Hfq hexamer. Cyan, Hfq65; orange, Hfq65-Q35A; pink, Hfq65-K47A; green, Hfq65-R19D; steel-blue, Hfq65-R16A. Rate constants are the average of 5 technical replicates with standard deviations less than 5%. The vertical dashed line indicates the Hfq concentration for which annealing progress curves are shown in Figure 5.3D and Panel B. The single turnover annealing rate reaches a maximum at equimolar concentrations of (Hfq)6:beacon. Higher Hfq concentrations can inhibit annealing due to random-order binding of RNA substrates and the formation of Hfq_{12}, which is inactive. (B) Target-A18 (distal) annealed by 33 nM Hfq65 hexamer and variants on Hfq65 background. The change in fluorescence emission intensity was normalized to the maximum fluorescence within an experiment. The average of five measurements is shown per progress curve. All progress curves were fitted to single- or double-exponential rate equations to obtain k_{obs}, as previously described[30].

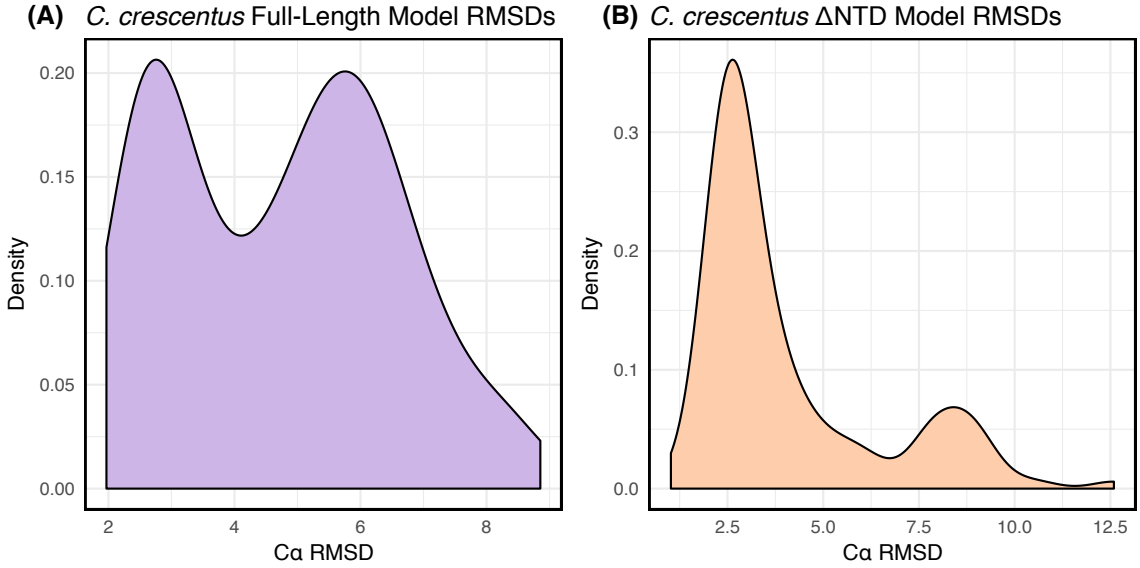

Figure 5.A.6: The alpha carbon coordinates of each subunit from Hfq hexamers modelled with Rosetta FloppyTail were compared to the coordinates of the subunit with the greatest number of resolved residues from the crystallographic structure. (A) Low-energy *C. crescentus* Hfq models have a significant population of structures with low-RMSD. (B) N-terminal extension residues were excluded from the simulation to eliminate NTD–CTD interactions. Excluding the NTD during modelling increases the low-RMSD population of models.

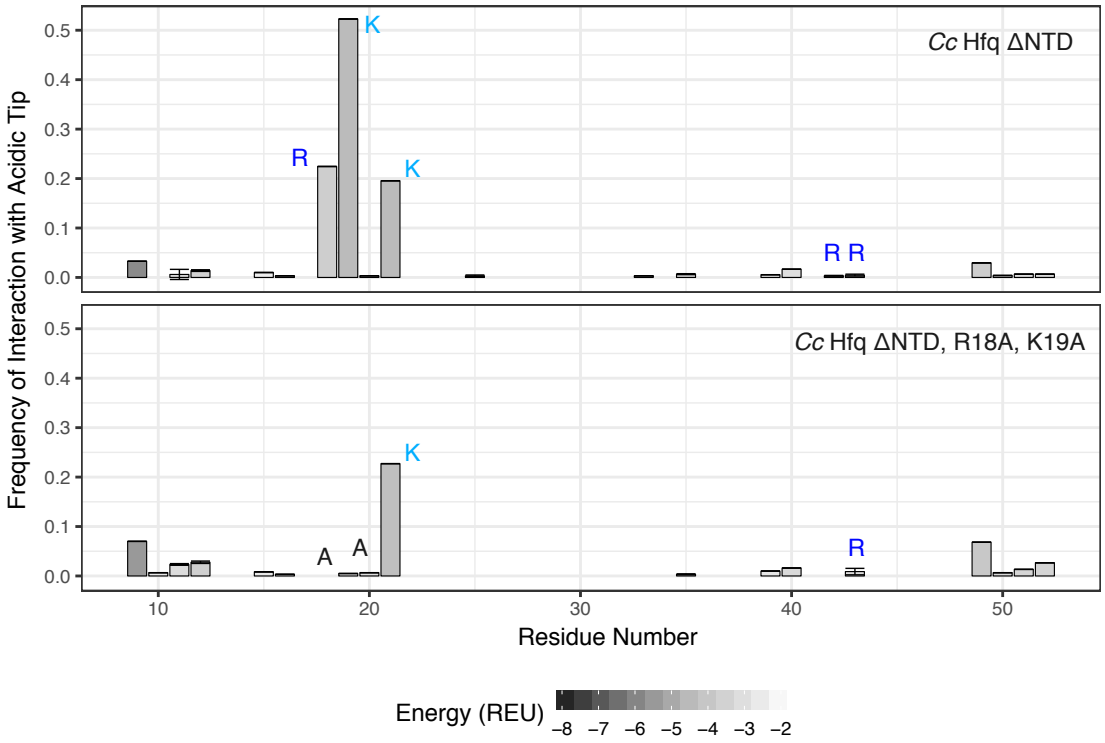

Figure 5.A.7: Observed frequency of favourable ($E < -2.0$ REU) core residue-to-acidic CTD interactions in low-energy FloppyTail models for Cc Hfq excluding the N-terminus (top panel) and a variant with two rim residues mutated to alanine (R18A, K19A; bottom panel). Mutating out the positive rim residues obviates interactions *in silico*. Error bars show $\pm$ one standard deviation, as computed by bootstrap resampling.

References

1. Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annual Review of Biochemistry* **83,** 553–584 (2014).
2. Oates, M. E. *et al.* D2P2: database of disordered protein predictions. *Nucleic Acids Research* **41,** D508–D516 (2012).
3. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. *Annual Review of Biophysics* **37,** 215–246 (2008).
4. Woodson, S. A., Panja, S. & Santiago-Frangos, A. in *Regulating with RNA in Bacteria and Archaea* 4, 385–397 (asm Pub2Web, 2018).
5. Saavedra, H. G., Wrabl, J. O., Anderson, J. A., Li, J. & Hilser, V. J. Dynamic allostery can drive cold adaptation in enzymes. *Nature* **558,** 324–328 (2018).
6. Alexander, E. J. *et al.* Ubiquilin 2 modulates ALS/FTD-linked FUS–RNA complex dynamics and stress granule formation. *Proceedings of the National Academy of Sciences* **115,** 201811997 (2018).
7. Ban, D., Iconaru, L. I., Ramanathan, A., Zuo, J. & Kriwacki, R. W. A small molecule causes a population shift in the conformational landscape of an intrinsically disordered protein. *Journal of the American Chemical Society* **139,** 13692–13700 (2017).
8. Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S. & Shaw, D. E. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society* **134,** 3787–3791 (2012).
9. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences* **107,** 8183–8188 (2010).
10. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences* **110,** 13392–13397 (2013).
11. Mittal, A., Lyle, N., Harmon, T. S. & Pappu, R. V. Hamiltonian Switch Metropolis Monte Carlo Simulations for Improved Conformational Sampling of Intrinsically Disordered Regions Tethered to Ordered Domains of Proteins. *Journal of Chemical Theory and Computation* **10,** 3550–3562 (2014).
12. Kleiger, G., Saha, A., Lewis, S., Kuhlman, B. & Deshaies, R. J. Rapid E2-E3 Assembly and Disassembly Enable Processive Ubiquitylation of Cullin-RING Ubiquitin Ligase Substrates. eng. *Cell* **139,** 957–968 (2009).

13. Crawley, S. W. *et al.* Autophosphorylation activates Dictyostelium myosin II heavy chain kinase A by providing a ligand for an allosteric binding site in the α-kinase domain. en. *Journal of Biological Chemistry* **286,** 2607–2616 (2011).

14. Zhang, J., Lewis, S. M., Kuhlman, B. & Lee, A. L. Supertertiary structure of the MAGUK core from PSD-95. *Structure* **21,** 402–413 (2013).

15. Zhang, A., Wassarman, K. M., Ortega, J., Steven, A. C. & Storz, G. The Sm-like Hfq Protein Increases OxyS RNA Interaction with Target mRNAs. *Molecular Cell* **9,** 11–22 (2002).

16. Schumacher, M. A., Pearson, R. F., Møller, T., Valentin-Hansen, P. & Brennan, R. G. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: A bacterial Sm-like protein. *EMBO Journal* **21,** 3546–3556 (2002).

17. Mikulecky, P. J. *et al.* Escherichia coli Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. *Nature Structural & Molecular Biology* **11,** 1206–1214 (2004).

18. Link, T. M., Valentin-Hansen, P. & Brennan, R. G. Structure of Escherichia coli Hfq bound to polyriboadenylate RNA. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 19292–7 (2009).

19. Soper, T. J., Doxzen, K. & Woodson, S. A. Major role for mRNA binding and restructuring in sRNA recruitment by Hfq. *RNA (New York, N.Y.)* **17,** 1544–50 (2011).

20. Schu, D. J., Zhang, A., Gottesman, S. & Storz, G. Alternative Hfq-sRNA interaction modes dictate alternative mRNA recognition. *The EMBO Journal* **34,** 2557–2573 (2015).

21. Małecka, E. M., Strózecka, J., Sobańska, D. & Olejniczak, M. Structure of bacterial regulatory RNAS determines their performance in competition for the chaperone protein HFQ. *Biochemistry* **54,** 1157–1170 (2015).

22. Otaka, H., Ishikawa, H., Morita, T. & Aiba, H. PolyU tail of rho-independent terminator of bacterial small RNAs is essential for Hfq action. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 13059–64 (2011).

23. Ishikawa, H., Otaka, H., Maki, K., Morita, T. & Aiba, H. The functional Hfq-binding module of bacterial sRNAs consists of a double or single hairpin preceded by a U-rich sequence and followed by a 3′ poly(U) tail. *RNA* **18,** 1062–74 (2012).

24. Zhang, A., Schu, D. J., Tjaden, B. C., Storz, G. & Gottesman, S. Mutations in Interaction Surfaces Differentially Impact E. coli Hfq Association with Small RNAs and Their mRNA Targets. *Journal of Molecular Biology* **425,** 3678–3697 (2013).

25. Panja, S. & Woodson, S. A. Hfq proximity and orientation controls RNA annealing. *Nucleic Acids Research* **40,** 8690–8697 (2012).

26. Zheng, A., Panja, S. & Woodson, S. A. Arginine Patch Predicts the RNA Annealing Activity of Hfq from Gram-Negative and Gram-Positive Bacteria. *Journal of Molecular Biology* **428,** 2259–2264 (2016).

27. Beich-Frandsen, M., Večerek, B., Sjöblom, B., Bläsi, U. & Djinović-Carugo, K. Structural analysis of full-length Hfq from Escherichia coli. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **67,** 536–540 (2011).

28. Vincent, H. A. *et al.* Characterization of Vibrio cholerae Hfq Provides Novel Insights into the Role of the Hfq C-Terminal Region. *Journal of Molecular Biology* **420,** 56–69 (2012).

29. Santiago-Frangos, A. & Woodson, S. A. Hfq chaperone brings speed dating to bacterial sRNA. *Wiley Interdisciplinary Reviews: RNA* **9** (2018).

30. Santiago-Frangos, A., Kavita, K., Schu, D. J., Gottesman, S. & Woodson, S. A. C-terminal domain of the RNA chaperone Hfq drives sRNA competition and release of target RNA. *Proceedings of the National Academy of Sciences of the United States of America* **113,** E6089–E6096 (2016).

31. Chaudhury, S., Lyskov, S. & Gray, J. J. *PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta* 2010.

32. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science* **23,** 47–55 (2014).

33. Nivon, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS ONE* **8** (ed Zhang, Y.) e59004 (2013).

34. O'Meara, M. J. *et al.* Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *Journal of Chemical Theory and Computation* **11,** 609–622 (2015).

35. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13,** 3031–3048 (2017).

36. Chaudhury, S. *et al.* Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS ONE* **6** (ed Uversky, V. N.) e22477 (2011).

37. Hopkins, J. F., Panja, S., McNeil, S. A. N. & Woodson, S. A. Effect of salt and RNA structure on annealing and strand displacement by Hfq. *Nucleic Acids Research* **37,** 6205–6213 (2009).

38. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research* **43,** 204–212 (2014).

39. Yamada, K. D., Tomii, K. & Katoh, K. Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. *Bioinformatics* **32,** 3246–3251 (2016).

40. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).

41. Gouveia-Oliveira, R., Sackett, P. W. & Pedersen, A. G. MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics* **8,** 312 (2007).

42. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A sequence logo generator. en. *Genome Research* **14,** 1188–1190 (2004).

43. Gottesman, S *et al.* Small RNA regulators and the bacterial response to stress. *Cold Spring Harbor symposia on quantitative biology* **71,** 1–11 (2006).

44. Beich-Frandsen, M. *et al.* Structural insights into the dynamics and function of the C-terminus of the E. coli RNA chaperone Hfq. en. *Nucleic Acids Research* **39,** 4900–4915 (2011).

45. Bohn, C., Rigoulay, C. & Bouloc, P. No detectable effect of RNA-binding protein Hfq absence in Staphylococcus aureus. *BMC Microbiology* **7,** 10 (2007).

46. Liu, Y. *et al.* Hfq Is a Global Regulator That Controls the Pathogenicity of Staphylococcus aureus. *PLoS ONE* **5** (ed Wang, P.) e13069 (2010).

47. Rochat, T. *et al.* Tracking the Elusive Function of Bacillus subtilis Hfq. *PLOS ONE* **10** (ed Randau, L.) e0124977 (2015).

48. Christiansen, J. K., Larsen, M. H., Ingmer, H., Søgaard-Andersen, L. & Kallipolitis, B. H. The RNA-binding protein Hfq of Listeria monocytogenes: role in stress tolerance and virulence. *Journal of bacteriology* **186,** 3355–62 (2004).

49. Oglesby-Sherrouse, A. G. & Vasil, M. L. Characterization of a Heme-Regulated Non-Coding RNA Encoded by the prrF Locus of Pseudomonas aeruginosa. *PLoS ONE* **5** (ed Rénia, L.) e9930 (2010).

50. Tsui, H.-C. T., Leung, H.-C. E. & Winkler, M. E. Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in Escherichia coli K-12. *Molecular Microbiology* **13,** 35–49 (1994).

51. Nielsen, J. S. *et al.* Defining a role for Hfq in Gram-positive bacteria: evidence for Hfq-dependent antisense regulation in Listeria monocytogenes. *Nucleic Acids Research* **38,** 907–919 (2010).

52. Rochat, T., Bouloc, P., Yang, Q., Bossi, L. & Figueroa-Bossi, N. Lack of interchangeability of Hfq-like proteins. *Biochimie* **94,** 1554–1559 (2012).

53. Trudeau, T. *et al.* Structure and Intrinsic Disorder in Protein Autoinhibition. *Structure* **21,** 332–341 (2013).

54. Varadi, M., Zsolyomi, F., Guharoy, M. & Tompa, P. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLOS ONE* **10** (ed Levy, Y. K.) e0139731 (2015).

55. Järvelin, A. I., Noerenberg, M., Davis, I. & Castello, A. The new (dis)order in RNA regulation. *Cell Communication and Signaling* **14,** 9 (2016).

56. Qualley, D. F. *et al.* C-terminal domain modulates the nucleic acid chaperone activity of human T-cell leukemia virus type 1 nucleocapsid protein via an electrostatic mechanism. *The Journal of biological chemistry* **285,** 295–307 (2010).

57. Tretter, E. M. & Berger, J. M. Mechanisms for defining supercoiling set point of DNA gyrase orthologs: I. A nonconserved acidic C-terminal tail modulates Escherichia coli gyrase activity. *The Journal of biological chemistry* **287,** 18636–44 (2012).

58. Kozlov, A. G., Cox, M. M. & Lohman, T. M. Regulation of Single-stranded DNA Binding by the C Termini of Escherichia coli Single-stranded DNA-binding (SSB) Protein. *Journal of Biological Chemistry* **285,** 17246–17252 (2010).

59. Watson, M., Stott, K. & Thomas, J. O. Mapping Intramolecular Interactions between Domains in HMGB1 using a Tail-truncation Approach. *Journal of Molecular Biology* **374,** 1286–1297 (2007).

60. Kobe, B. & Kemp, B. E. Active site-directed protein regulation. *Nature* **402,** 373–376 (1999).

Chapter 6

Re-Design of Protein Crystals

6.1 Overview

Substantial advances have been made in the computational design of protein interfaces over the last 20 years. However, the interfaces targeted by design have typically been stable and high affinity. Here, I report the development of a generic computational design method to stabilize the tenuous interactions at crystallographic interfaces. Initially, I analyzed structures reported in the Protein Data Bank (PDB) to determine whether crystals with more stable interfaces result in higher resolution structures. I found that, for twenty-two variants of a single protein, crystallized by a single individual, Rosetta score correlates with resolution. I then developed and tested a computational design protocol, seeking to identify point mutations that would improve resolution, on a highly stable variant of staphylococcal nuclease (SNase Δ+PHS). Only one of eleven initial designs crystallized, forcing me to re-evaluate my design strategy and base my designs on an ensemble of protein backbones. Using this approach, four of the seven designs crystallized. Collecting diffraction data for multiple crystals per design and solving crystal structures, I found

that designed crystals improved resolution modestly and in unpredictable ways, including altering crystal space group. *Post-hoc*, *in silico* analysis showed that crystal space groups could have been predicted for four of six variants (including WT), but that resolution did not correlate with interface stability, as it did in the preliminary results. My results show that single point mutations can have significant effects on crystal resolution and space group, and that it is possible to computationally identify such mutations, suggesting a potential design strategy to generate high-resolution protein crystals from poorly diffracting ones.

6.2 Introduction

X-ray crystallography is still the primary method for acquiring atomic-scale structural information about biological macromolecules such as proteins, and it is indispensable for gaining functional and mechanistic insights across biological and pharmacological disciplines[1]. However, because of its highly unpredictable nature, crystallography is viewed more often as art than as science–a fact reflected by the low rate of success in large-scale protein crystallization efforts ($\sim$10–20%)[2,3]. Sometimes, when proteins produce diffraction-quality crystals, the data may be of low quality or unsolvable. Even when a crystal structure can be determined, some regions might be missing. For example, approximately 23% of the crystal structures reported to the PDB diffract to a resolution of $\geq$2.5 Å[4]. At a resolution of 2.5 Å, the backbone, side chains, and small molecules can be fit with a reasonable degree of precision to the electron density; however, key features such as the placement of water molecules or alternate side-chain conformations may be less certain. At even lower resolutions (3–6 Å), ligands or side chains and even the main chain may not be fit reliably[5–8]. An inability to resolve ligands, water molecules, small molecules, or side-chain interactions prevents accurate understanding of catalytic mechanisms, drug-protein interactions, or the organization of certain macromolecular complexes, and precludes computational design from using natural proteins as input.

Historically, rational design has been used to overcome various degrees of protein recalcitrance to crystallization, from improving existing crystals to generating new ones[9]. The variety in strategies has been quite broad. Some strategies can be applied when only the protein sequence is known, even before crystal trays are laid, *e.g.* deleting loops or regions of low sequence complexity[10], or by identifying stabilizing mutations from homologous sequences[11], surface entropy reduction (SER)[12], or *de novo* crystal design[13]. Other strategies have focused on improving an existing crystal, such as through the rational engineering of crystal contacts[14,15]. Of the above strategies, all but SER must be tailored to a specific target protein. The necessity for protein-specific approaches is somewhat surprising considering that the underlying physics is universal. For example, Fusco *et al.* identified two generic mechanisms underlying crystal formation in their analysis of 182 proteins in 1,536 crystallization conditions[16]. In principle, a reliable and general method for enhancing the resolution of poorly-diffracting crystals through rational and computational design should exist.

Here, I report my attempts to develop resolution-enhancing computational design of protein–protein interactions at crystallographic interfaces. I began by identifying for the physical determinants of high-resolution protein crystals. This lead me to identify a positive correlation between resolution and crystal lattice stability (the Rosetta-determined score of the asymmetric unit and the unique protein–protein interactions defining the crystal lattice). A Rosetta protocol was then developed to identify stabilizing (resolution-enhancing) point mutations. The protocol was benchmarked *in silico* against rationally-engineered protein crystals[17]. I tested my protocol experimentally by designing, cloning, expressing, purifying, and crystallizing variants of staphylococcal nuclease (SNase). I found that variants designed on a single, fixed backbone crystallized rarely (1/11), whereas variants designed on an ensemble of backbones crystallized more readily (4/7). Comparison of the highest-resolution shells for collected diffraction data (determined by $CC_{1/2}$) revealed only minor improvements in resolution ($\sim$0.05 Å) for three of five designs that crystallized.

Surprisingly, two of the resolution-enhancing designs altered the crystal space group. An analysis of my efforts shows that space group changes could have been predicted for three of the five designs, but that crystal lattice stability does not correlate with resolution for my test protein.

6.3 Methods

All data associated with this chapter are available online via Zenodo at the following DOIs:

- 10.5281/zenodo.3216968
- 10.5281/zenodo.3222946
- 10.5281/zenodo.3228344
- 10.5281/zenodo.3228838
- 10.5281/zenodo.3235486
- 10.5281/zenodo.3235518
- 10.5281/zenodo.3228351

6.3.1 Curation of crystal datasets

Three protein structure datasets were constructed on October 10th, 2016, from the Protein Data Bank[4] (PDB), termed the: "PDB representative", "SNase", and "Mizutani" sets.

To generate the PDB representative set, I first generated three lists of PDB IDs and then took the intersection of the lists. The first list ensured that I only analyzed non-redundant, reasonable-quality protein structures. Using the PISCES server[18], I generated a list of non-Cα-only X-ray structures in the PDB adhering to the following criteria (culling by chain):

- 25% maximum sequence identity,
- resolution better than 3.0 Å,
- R-value < 0.3, and

- proteins comprising more than 40, but less than 10,000 residues.

This list, `pisces.txt`, contained 10,886 PDB IDs. Next, I generated a second list to limit my analysis to solely crystallographic protein–protein interactions. Using the advanced search option on the PDB website, I generated a list of PDB structures with monomeric stoichiometry and only a single chain in both the biological and asymmetric units. This list, `pdb.txt`, contained 36,899 PDB IDs. Finally, I generated a third list to exclude structures containing many ligands or non-protein atoms. Starting with the PDB IDs from the `pisces.txt` list, I used a Python script (`1-parse-pdb-remarks.py`) to filter PDB IDs, selecting for the absence of `REMARK` 465/470/475/480 records, which indicate missing atoms, and a fraction of non-`HOH HETATM` records greater than 0.1. This list, `missing_or_nonhet.txt`, contained 873 PDB IDs. I took the intersection of my three lists of PDB IDs as my "PDB representative" set; this was performed in R using the `merge-three-lists.R script`. The final list contained 379 PDB IDs.

Separately, to generate the SNase set, I used a new PDB advanced search to select for X-ray structures with UniProt Accession ID: P00644, in addition to the above criteria for stoichiometry, biological unit, and asymmetric unit, but not culling for sequence identity, R-value less than 0.3, absence of atoms or presence of ligands. The SNase set contained 256 PDB IDs, which were not present in the PDB representative set.

Finally, the Mizutani set was simply composed of the 21 structures of diphtine synthase deposited by Mizutani *et al.* in their study of rational crystal contact engineering[17]. These structures were not present in either the PDB representative or SNase set.

6.3.2 Modeling of crystals

In this study, I sought to computationally quantify crystallographic protein–protein interactions. To that end, proteins were modeled in three states: (1) as a crystal, including all symmetry mates within 12 Å, (2) as a collection of pairwise interfaces, and (3) as a monomer. These states were constructed for each dataset. Furthermore, pairwise interfaces were

analyzed with the Evolutionary Protein–Protein Interface Classifier[19] (EPPIC) to verify that the interactions I was assessing were crystallographic and not biological.

PDB Representative: Monomers were downloaded from the PDB and energy minimized using Rosetta, weekly version: v2018.24. To ensure accurate energy calculation and because Rosetta cannot model all ligands/co-factors/etc., `HETATM` records were omitted. Energy minimization was performed by the FastRelax protocol[20,21] with the following command line:

```
relax.linuxgccrelease -l list.txt -relax:ramp_constraints -relax:
   constrain_relax_to_start_coords -ex1 -ex2 -use_input_sc -flip_NHQ -
   no_optH false -nstruct 10 -out:pdb_gz
```

where `list.txt` contained the PDBs. Individual crystallographic interfaces were generated and analyzed using the pre-compiled EPPIC command-line interface (version 3.0.5):

```
epicc -i 1ABC.pdb -l -p -s
```

where 1ABC.pdb is any PDB. Any PDB ID with an interface predicted by EPPIC to be biological and not crystallographic was excluded from further analysis if the corresponding PDB entry or supporting literature indicated that the biologically relevant state was not monomeric (as I only wished to study crystallographic interactions). Crystals were modeled using the same protocol, except this time the `-symmetry:symmetry_definition CRYST1` flag was included to enable modeling of the asymmetric unit and all symmetry mates within 12 Å as previously described[22]. The energy of crystallization was computationally determined, using the weekly PyRosetta[23] release, v2018.24, and the November 2016 version of the Rosetta scoring function[24,25]. The script `score_crystal_interfaces_parallel.py` evaluated the energy of each crystallographic interface, which was later combined with the energy of the monomer to yield the crystal energy (see Results). After excluding four structures that could not be modeled with my approach and twelve structures that accidentally included biological interactions in the crystal, 364 PDBs were analyzed from the PDB. These PDBs are listed in: `pdb-representative-rosetta.txt`.

SNase: As above, SNase monomers were downloaded from the PDB and energy min-

imized using Rosetta v2018.24. Unlike above, HETATM records were retained, because the nucleotide analog thymidine-3′,5′-diphosphate (THP) bound by the enzyme makes important crystal contacts. The ligand geometry was fixed in simulations and read from the PDB Chemical Components Dictionary. Energy minimization was performed as described above. Individual crystallographic interfaces were not generated and analyzed using EPPIC because SNase is known to be a monomer. Crystals were modeled and crystal energies were computed as described above.

Mizutani: Modeling of the 21 diphtine synthase structures from Mizutani *et al.* was performed similarly to modeling of the PDB and SNase crystal structures, as described above. The only exception being that diphtine synthase is naturally a dimer, so during energy evaluation the two dimeric chains were treated as a monomer and only crystallographic, not biological, interfaces were considered. As with SNase, because the biological state is known, EPPIC was not used to generate or analyze crystallographic interfaces. Crystals were modeled and crystal energies were computed as described above.

6.3.3 Forward design of dipthine synthase

To determine the computational design approach most likely to be experimentally successful, I tested several strategies on diphthine synthase. Since the resolution is reported for twenty variants, I asked if a particular design strategy can predict resolution-improving variants. This approach is known as forward design. The energy-minimized structure of wild-type diphthine synthase (PDB ID: 1WNG) was used as input for design. At each position mutated by Mizutani *et al.* (26, 49, 54, 65, 69, 79, 140, 142, 146, 171, 173, 187, and 261), each amino acid except cysteine or proline was tested. I attempted to accommodate the mutation by permitting varying degrees of freedom in the wild-type crystal form (simulated by Rosetta Symmetry). These different design strategies ranged from only permitting neighboring side chains to repack to re-docking in the crystal lattice (see Results). All energy calculations were performed in the crystal form and averaged over ten repeats.

6.3.4 Computational design of SNase

Design for SNase variants initially followed my most successful diphthine synthase approach: introducing a point mutation at a position followed by repacking of side-chains followed by energy minimization of side-chain and backbone dihedral angles. An energy-minimized structure of Δ+PHS SNase (PDB ID: 3BDC) was used as input. The geometry of the THP ligand was held fixed during the simulation and defined by a Rosetta params file derived from the PDB coordinates. Later, I introduced an ensemble of 200 perturbed backbones generated by Rosetta Backrub[26] as it has been shown that interface ΔΔG prediction is more accurate when using an ensemble of backbones rather than just a single input[27]. The following steps were repeated for each member of the backbone ensemble and for every designable surface position, defined as residues having a Cα–Cα distance under 8 Å across any crystallographic interface. First, the position was mutated to one of eighteen amino acids (cysteine and proline were excluded). Then, the crystal form was generated using Rosetta Symmetry in the wild-type space group and unit cell dimensions. Finally, side-chains were repacked to accommodate the mutation in the crystal form. The error in this modeling was calculated across the ensemble of 200 backbones, instead of by repeating the simulation ten times.

6.3.5 Cloning, expression, and purification of proteins

Point mutations were introduced by Quikchange mutagenesis[28] into the highly stable Δ+PHS variant[29], expressed in *E. coli* BL21/DE3 cells transformed with the pET-24a+, and purified as previously described[30].

6.3.6 Protein Crystallization

Crystals of Δ+PHS and its variants were grown by the hanging drop vapor-diffusion method at 277 K. The reservoir solution varied, ranging in pH from 6–9, with 20–40% (v/v) 2-methyl-2,4-pentanediol (MPD), either 3 or 2 molar equivalents of THP, either 2 or 1 molar

equivalents $CaCl_2$, and 25 mM potassium phosphate. The protein concentration varied across variants, but was always mixed in a 1:1 ratio with the reservoir solution to make the drop. Conditions are detailed for each crystal in Supplemental Table 6.A.1. Crystals typically appeared after one week, were harvested with Hampton Research CryoLoops™ on CrystalCap™ Copper HT magnetic sample mounts, and were immediately flash-cooled in liquid nitrogen. Crystals were stored at 77 K until data collection.

6.3.7 Data collection and structure determination

X-ray diffraction data were collected for single crystals at 77 K using a Rigaku FR-E SuperBright rotating anode X-ray generator and a Rigaku DECTRIS PILATUS 200K pixel array detector. Diffraction data were indexed, integrated, and scaled using the XDS program package[31]. Phasing, modeling building and model refining was performed using PHENIX[32]. Phasing was performed by molecular replacement in PHASER[33] using the search model 3BDC, with solvent-exposed or mutated side chains truncated to the Cα position to avoid biasing side-chain placement at crystal contact sites. Side chains were rebuilt using COOT[34]: initial placement was performed using the Mutate and Autofit function and followed by manual refinement. Whole-structure refinement and water placement was performed using `phenix.refine`[35] and `phenix.rosetta_refine`[36]. Data collection and refinement statistics are shown in Supplemental Table 6.A.2. Crystal structures deposited to the PDB include models 6OK8 (K127L), 6OK9 (K133M), and 6OKA (Q123D).

6.4 Results

The primary goal in this effort was to develop a broadly applicable, Rosetta-based computational method for crystal contact design, with the goal of systematically predicting single point mutations that could enhance resolution. To this end, I first asked whether or not Rosetta scoring functions might correlate with resolution.

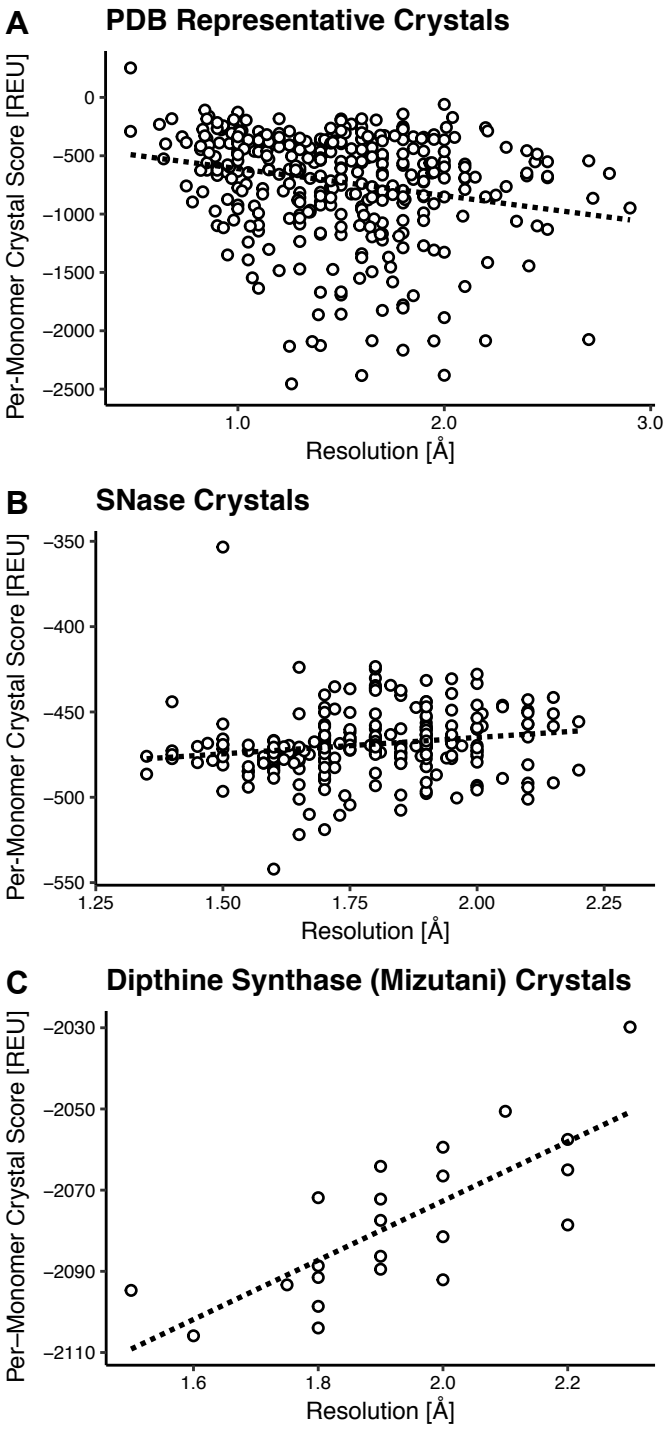

Figure 6.1: Caption follows on the next page.

Figure 6.1: (Continued from previous page.) Rosetta score correlates with resolution when comparing crystals of the same protein varying only by point mutations. The per-monomer crystal scores (*i.e.* the score of the monomer plus all crystallographic interactions, which is the minimal interacting unit required to generate the crystal) and crystal resolutions are compared for three sets of protein structures. The PDB-representative set (A) samples 364 monomers with distinct sequences and shows no relationship between score and resolution. The SNase set (B, excluding an outlier at 2.5 Å) compares only crystals of *Staphylococcal aureus* nuclease variants, attempting to rule out the protein as a variable, but still there is no trend. Finally, the dipthine synthase set (C) compares crystals that vary only by a point mutation, ruling out most extrinsic variables, and score correlates with resolution.

6.4.1 Rosetta score correlates with resolution, when other variables are controlled

With over 129,000 crystal structures of biological macromolecules[4], the PDB provides a trove of data that can be used to determine whether or not Rosetta score correlates with the resolution of protein crystals. To ensure a fair comparison, structures must be first energy-minimized in the Rosetta scoring function using the Rosetta FastRelax protocol[20,21]. As FastRelax runtime scales with protein size, testing every structure in the PDB is not feasible. Furthermore, some structures are overrepresented in the PDB, which might bias analyses. Instead of analyzing all structures, I selected a diverse and representative subset of the PDB containing 364 structures, which had a maximum sequence identity of 25%, a resolution better than 3 Å, R-values less than 0.3, and featured only crystallographic interactions (fully described in Methods). For every structure in the set, I generated all symmetry-mates within 12 Å using Rosetta Symmetry[22] and energy minimized this "crystal" form ten separate times using the default FastRelax protocol, which features four cycles of minimization each with progressively weaker harmonic constraints to prevent substantial deviation from the starting coordinates. Separately, I energy minimized the monomeric form of the protein, which was also the asymmetric unit and the biologically-relevant unit. I approximated the energy of the crystal as $E_C = \langle E_m \rangle + \frac{1}{2} \sum_i \langle E_i \rangle$, where $\langle E_m \rangle$ is the average Rosetta score of ten energy-minimized monomers and $\sum_i \langle E_i \rangle$ is the average Rosetta score of interface i in the crystal form, summed over all interfaces within 12 Å of the asymmetric unit. Hence, E_C represents that energy of the minimal unit required to generate the crystal.

The relationship between resolution and score is shown in Figure 6.1.

I initially found a slight anti-correlation between resolution and score for the representative PDB set: low-resolution structures had lower scores than high-resolution ones. I hypothesized that this unexpected result was caused by my inability to control for the many variables that affect resolution that are not captured by Rosetta score (*e.g.*, how the highest resolution-shell cutoff was decided, user handling, the content of the reservoir solution, *etc.*). To test this hypothesis, I analyzed two additional sets of crystal structures in the same manner as the PDB representative set. The first additional set I analyzed controlled for the protein as a variable. I searched for a small, globular protein, with many structures in the PDB that differed only slightly from each other, but that spanned at least 1 Å in resolution. Of the multiple proteins fulfilling these criteria, I selected SNase, which had 256 crystal structures. I used Rosetta Symmetry and FastRelax to generate ten energy-minimized monomers and crystals, and computed the energy of the crystal as described above. The SNase set did not show a strong correlation between resolution and score (Figure 1B). To further control extrinsic variables, I analyzed twenty-two variants of a single protein (diphthine synthase) that had been previously cloned, expressed, purified, crystallized and the crystal structures solved by one scientist in a crystal engineering study by Mizutani *et al.*[17]. Despite the variants only differing by a point mutation or two, the crystal structures spanned a range of resolutions from 1.5 Å to 2.3 Å. The structures were energy minimized and the crystal energies calculated in a similar fashion as for the previous two sets. The Mizutani set showed a strong correlation ($R^2 = 0.8$) between Rosetta score and resolution (Figure 6.1).

6.4.2 Rosetta can identify resolution-enhancing mutations

Since low score corresponded to high resolution for the Mizutani set, I next sought a fast computational design strategy that could identify resolution-enhancing mutations from the wildtype (WT) crystal structure. I tested six strategies on the Mizutani set in an approach

known as forward design. Using the energy-minimized crystal form of the wild-type protein as input, I introduced point mutations one-by-one at the positions engineered by Mizutani *et al.* I then optimized side-chain dihedral angles while keeping the backbone fixed (side-chain repacking)[37]. Following repacking, I tested six design strategies with varying degrees of freedom: (1) I did nothing else before evaluating the score, (2) I applied gradient-based energy minimization on side-chain dihedral angles, (3) I applied gradient-based energy minimization on side-chain and backbone dihedral angles, (4) I applied gradient-based energy minimization on side-chain dihedral angles and the relative position/orientation of the protein and its symmetry mates, (5) I applied gradient-based energy minimization on side-chain and backbone dihedral angles and the relative the relative position/orientation of the protein and its symmetry mates, and (6) I sampled the relative position/orientation of the protein and its symmetry mates, translating in steps of 0.05 Å and rotating in steps of 0.1 degrees, followed by energy minimization as in Strategy 5 over four Monte Carlo cycles. All gradient-based minimization was run until convergence was achieved, defined as a change in Rosetta score of less than 0.00001 following an iteration of minimization, or for 200 iterations. Each strategy was tested ten times to assess error. The forward design results for all strategies are shown in Supplemental Figure 6.A.1.

I found Strategy 3 (minimizing on side-chain and backbone torsion angles after repacking) to be the most successful. Figure 6.2 compares the predicted change in score between each variant and WT diphthine synthase for Strategy 3. This approach successfully predicts 6 of 17 (35.3%) resolution-enhancing mutations identified in the paper. In addition to these six, this approach predicts 46 other mutations to have lower energy than WT and thus could be potentially resolution-enhancing; however, these were not experimentally characterized by Mizutani *et al.*, so it is unclear if these predictions are correct. Assuming a worst-case scenario where these uncharacterized point mutations do not enhance resolution-enhancing mutations, this design approach would predict 6 resolution-enhancing mutations out of 52 possibilities or 11.5%. Both success rates, 35.3% and 11.5%, compare favorably with

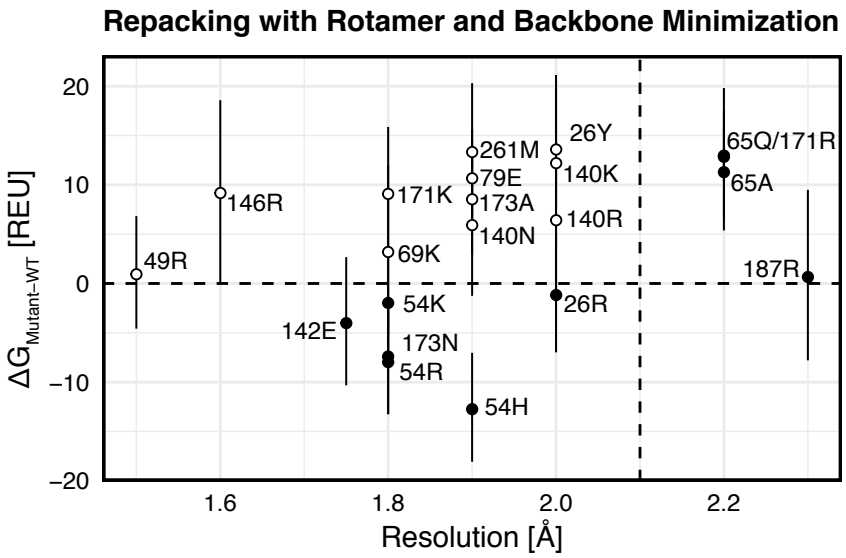

Figure 6.2: Forward design on dipthine synthase suggests that Rosetta can successfully identify mutations that improve or worsen resolution. The figure plots the difference in score (in the crystal form) between WT and various designs versus the experimentally determined resolution[17], with the dashed lines indicating the WT values. The black points are mutations whose score correctly predicts the sign of the resolution change (*i.e.* better than WT score results in better than WT resolution and *vice versa*) whereas the hollow points are mutations whose score incorrectly predicts resolution change. Standard deviations in score are calculated from 10 repeats of the design simulation.

historical protein interface design success rates, which are typically under 10%[38].

6.4.3 Rosetta-designed crystals slightly improve resolution

To determine whether my design approach was applicable to other proteins, I tested it on a model system: Δ+PHS, a user-friendly, highly stable variant of staphylococcal nuclease (SNase)[29]. I identified candidate designable residues at crystallographic interfaces as those with a Cα–Cα distance under 8 Å to neighboring symmetry mates. At each position, I introduced a point mutation followed by side-chain repacking and energy minimization of side-chain and backbone dihedral angles. I selected eleven designs for experimental characterization. However, of the eleven, only a single variant crystallized in conditions where the WT protein normally crystallizes. Since I wanted my approach to yield crystals without having to reoptimize crystal growth conditions, I sought to improve the crystallization rate of my designs. To this end, I introduced a step to generate backbone diversity before design. I drew inspiration from recent work showing that interfacial ΔΔG

calculations are more accurate when the change in energy is computed across an ensemble of models[27]. Since backbone diversity was introduced beforehand, I was more conservative in my approach and I followed the introduction of point mutations with only side-chain repacking (Strategy 1). From the second round of design, I identified seven possible variants, but since two overlapped with those found in the first round, only the five new variants were experimentally characterized. With this design approach, four of the seven variants yielded crystals in WT-like conditions; thus, designing on an ensemble of structures had improved my crystallization rate from 9% to 57%.

Next, I determined the resolution of the diffraction data collected for my variants and compared it to that of the WT protein. To control for differences across crystals, I collected full diffraction data sets for at least three crystals of each variant (up to a maximum of fifteen), depending on the propensity of each variant to form diffraction-quality crystals. The K127L variant, in particular, affected crystal growth and nucleation significantly, yielding larger crystals across more conditions than the other variants. I then indexed, integrated, and scaled the data sets using XDS. The most likely space group was determined by POINTLESS[39]. For all variants except K64R and K127L, this was the WT space group ($P2_1$). I found that K64R crystallized in $P2_12_12_1$ and K127L crystallized in $P4_1$, the third and second most common space groups for SNase crystals, respectively. In total, I were able to process the data for 37 of the 43 crystal diffraction patterns collected, with the remaining six datasets failing to index due to issues such as ice rings or poor spot profiles. I report a summary of the collected and processed diffraction data in Supplemental Table 6.A.1.

Following processing with XDS, I identified the highest resolution shell as the shell with the highest resolution still having a significant $CC_{1/2}$ ($t < 0.01$). I used $CC_{1/2}$, or the correlation between intensities when the data is split in half, to select the highest resolution shell because it provides a rigorous statistical cutoff[40]. To compute significance I calculated a t-value, $t = r\sqrt{\frac{n-2}{1-r^2}}$, where r is the $CC_{1/2}$ value and n is the number of reflection pairs (the degrees of freedom), and compared it to Student's t-distribution with the same degrees

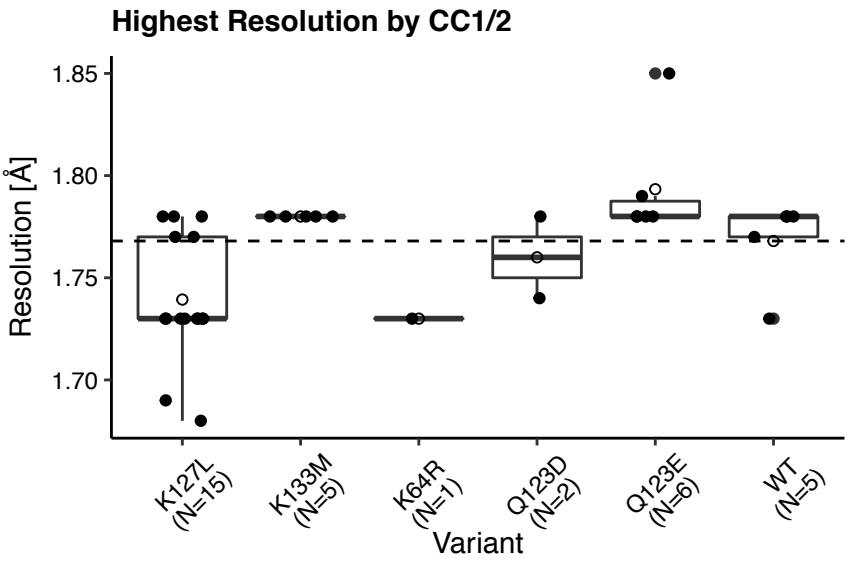

Figure 6.3: Distributions of the highest resolution shell show that some designs improve on WT resolution. Data were collected from multiple X-ray diffraction experiments and determined by significant $CC_{1/2}$ according to Student's t-test. Boxplots show the median resolution $\pm$ one quartile. Open circles indicate the average resolution. The dashed line is the average WT resolution. Designs K127L, K64R, and Q123D have higher average resolution than WT.

of freedom[41]. I found that, on average, three of the five designs (60%: Q123D, K64R, and K127L) achieved a higher resolution than WT (Figure 6.3). However, the improvement was minimal (<0.05 Å). In general, variant resolutions fell within a very narrow range: 1.67–1.85 Å, the width of which was only slightly greater than the range typically spanned by the resolutions of multiple crystals from the same variant ($\sim$0.1 Å).

6.4.4 Rosetta-designed crystals do not behave as predicted

Intrigued by the unexpectedly small variation in resolution between designed variants, I solved the crystal structures of several candidates to ask whether there was an underlying structural basis for the changes in resolution. I discuss the variants below, grouped by their observed effects on SNase crystallization, and provide a general summary of observations across all variants.

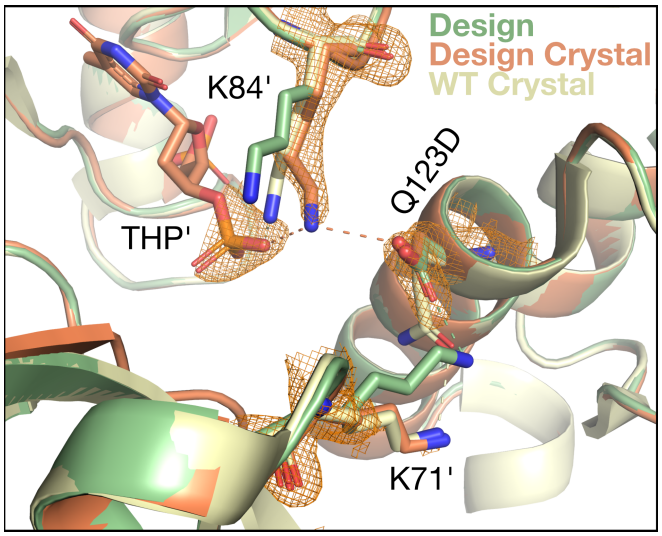

Figure 6.4: Q123D design (green) predicts the correct interaction type, but the incorrect interaction residue. In the design, residue D123 is predicted to form an electrostatic interaction with residue K71′ (′ indicates symmetry mate), improving on the WT Q–K interaction (pale yellow). However, this interaction is missing in the density and crystal structure of the variant (both orange; the 2mFo-DFc map contoured at 1.5σ for the Q123D variant crystal structure is carved within 2 Å of residues 71, 84, and 123). In place of the Q123–K71 interaction, D123 hydrogen bonds to K84, which also non-covalently interacts with the nucleotide analog (thymidine-3′,5′-diphosphate, THP) bound in the SNase active site. In this figure, each residue belongs to either the asymmetric unit or a different symmetry mate. Key interactions with atom-pair distances under 3.5 Å are shown as dashed lines.

6.4.4.1 Q123D and Q123E

In silico, the Q123D design strengthened the crystallographic interface by introducing an electrostatic interaction between D123 in the asymmetric unit and K71 in a neighboring symmetry mate. Upon solving the crystal structure, I found minor changes (less than 0.5 Å RMSD) in the backbone conformation (Supplemental Figure 6.A.2). Analysis of the site around the Q123D mutation revealed that residue D123 interacted with residue K84 of a neighboring symmetry mate (with a 3.1 Å distance between the lysine nitrogen and aspartic acid oxygen), instead of K71 (Figure 6.4). This result is in contrast to the WT structure, where residue K71 interacts with Q123 (3 Å distance between the corresponding oxygen and nitrogen atoms) and K84 solely interacts with the phosphate oxygen of THP (Supplemental Figure 6.A.3).

Although I was unable to solve the crystal structure for the Q123E variant due to twinning and the presence of ice, I expect a similar interaction to be occurring. This

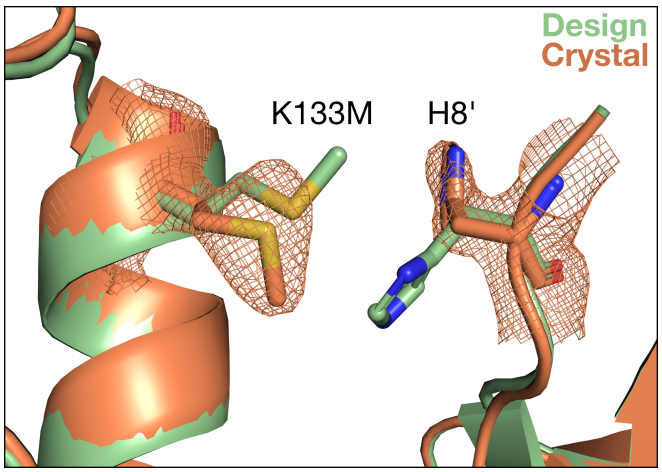

Figure 6.5: Superposition of the designed (green) and crystallized (orange) K133M structure. The 2mFo-DFc map is contoured at 1.5σ and carved within 2 Å of residues 8 and 133. The designed packing interaction does not occur in crystal, instead the side chains occupy alternative rotamers.

supposition is supported by the observed average resolutions, which are quite similar to the WT protein: Q123D slightly improves (by 0.01 Å) resolution whereas Q123E slightly worsens resolution (by 0.02 Å).

6.4.4.2 K133M

Like Q123E and Q123D, the K133M variant did not significantly alter resolution with respect to the WT protein and resulted in minimal backbone movements (0.17 Å backbone RMSD, Supplemental Figure 6.A.2). The design was favored *in silico* because it replaced an unfavorable electrostatic interaction between K133 and H8 with a van der Waals contact between M133 and H8, while also slightly reducing the entropic cost of forming that crystal contact (Figure 6.5). However, the K133M crystal structure revealed that, although the interface had compacted slightly, the side chains were too distant to interact. Compared to the design, the minimum distance between M133 and H8 in the crystal grew from 3.6 Å to 5.3 Å. This lack of interacting side chains likely explains the minimal effect of this mutation on crystal resolution.

6.4.4.3 K64R and K127L

The variant K127L produced the highest-resolution crystals in my study, though it did so in a manner not predicted by design: it altered the crystal space group. The WT protein crystallizes in the space group $P2_1$, and, in this crystal form, K127 forms a salt bridge with the THP molecule bound in the neighboring SNase active site. When this lysine is mutated to leucine, the salt-bridge interaction cannot form, destabilizing the $P2_1$ crystal form. Instead, the designed protein crystallizes in $P4_1$, a higher symmetry space group, where L127 packs against a neighboring loop by forming backbone interactions with K28 and G29. In this space group, K71 replaces K127 as the interacting partner of the THP in the crystal form, suggesting that the interaction of the substrate phosphate groups with a positively-charged side chain might be useful for SNase crystallization (Figure 6.6).

In addition to the space group change, the K127L variant had the largest backbone motions of all variants. These motions are in the loop region (residues 114–118) that precedes the α-helix containing L127. They occur when residue K116 shifts from contacting the neighboring molecule in the crystal to make contacts to the bound nucleotide analog instead.

The other variant that improved resolution in my study, albeit with a sample size of one, was K64R. Although I was unable to solve a crystal structure due to problems with twinning and the presence of ice, I was able to determine from the diffraction data that K64R (like K127L) resulted in a change in space group, going from $P2_1$ to the higher symmetry space group $P2_12_12_1$. To gain structural insight as to why this variant and space group might lead to higher resolution, I aligned my K64R model to a different SNase structure crystallized in the same space group (PDB ID: 5KEE) and applied the symmetry operations necessary to generate the neighboring symmetry mates, using unit cell dimensions from the diffraction data. I then used Rosetta's FastRelax protocol to alleviate any clashes that might have been introduced. I observed two possible hydrogen bonding interactions for

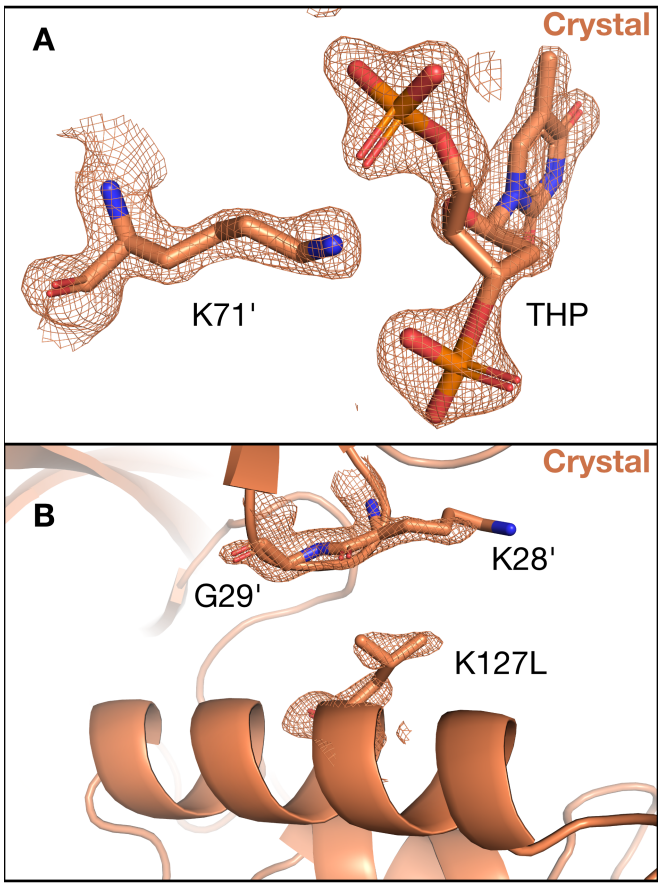

Figure 6.6: The variant, K127L, which yields the highest-resolution crystals, crystallizes in a higher symmetry space group ($P4_1$) than WT ($P2_1$) because it breaks an electrostatic contact central to a crystallographic interface in $P2_1$. (A) In the K127L crystal, the previous K127–THP salt bridge was retained, albeit with a different lysine residue (71). The 2mFo-DFc map is contoured at 1.5σ and carved within 2 Å of residues 71 and the THP molecule. (B) The new crystallographic interface containing L127 is well-resolved in density, and features non-specific side-chain–backbone contacts. The 2mFo-DFc map is contoured at 1.5σ and carved within 2 Å of residues 28, 29 and 127.

R64 that might account for this change in space group: one with the carboxylic acid of E135 of a neighboring symmetry mate and another with the backbone carbonyl of the oxygen of the same residue (Figure 6.7).

6.4.5 Retrospectively: Rosetta score recovers space group changes, but not resolution

Since I did not include the possibility of space group changes in my design protocol, yet I observed changes for two variants, I retrospectively asked whether Rosetta could recover the correct space group (Figure 6.8) by modeling and scoring each variant and the WT in each of the three most popular SNase space groups. For four out of six crystals (including WT), I found that Rosetta could correctly predict the space groups. Rosetta failed to predict the correct space group changes for the K127L variant, yielding $P2_12_12_1$ as the lowest scoring space group (when the actual space group was $P4_1$), and for the K64R substitution, yielding $P2_1$ as the lowest scoring space group (the actual space group was $P2_12_12_1$).

Finally, I asked whether the Rosetta score of the solved crystal structures correlated with the resolution, as I found to be the case for the engineered variants studied by Mizutani *et al.*[17]. I analyzed the crystal structures of my designs as previously described. Surprisingly, I found an anti-correlation between score and resolution (Figure 6.9), although it should be noted that this resolution range only spans ∼0.1 Å, whereas ∼0.8 Å was spanned by the crystal structures from Mizutani *et al.* (Figure 6.1).

6.5 Discussion

I attempted to develop and validate a generic computational method for protein crystal contact design to engineer crystals that yield high-resolution structural information. Probing the PDB, I found that Rosetta score correlated with crystal resolution when accounting for common external variables relevant to crystallization. Using data from an existing study in crystal engineering[17], I developed a design approach that recapitulated resolution-

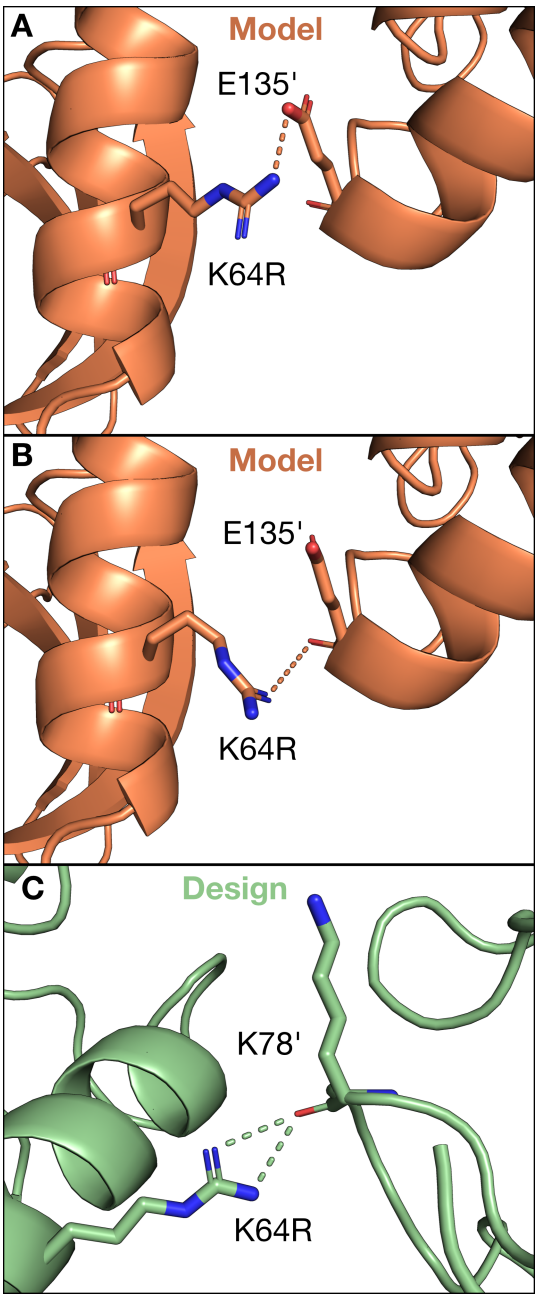

Figure 6.7: Models (orange) of the possible K64R interactions in the $P2_12_12_1$ space group show two new possible electrostatic interactions, with either the side-chain or backbone atoms of glutamic acid 135 in the neighboring symmetry mate (indicated by the ′). Residue K64 was not strongly interacting in WT, showing multiple possible rotameric states in electron density and missing density for some atoms (Supplemental Figure 6.A.4). Thus, the introduced R64–K78′ interaction in the design (green) was intended to be stabilizing.

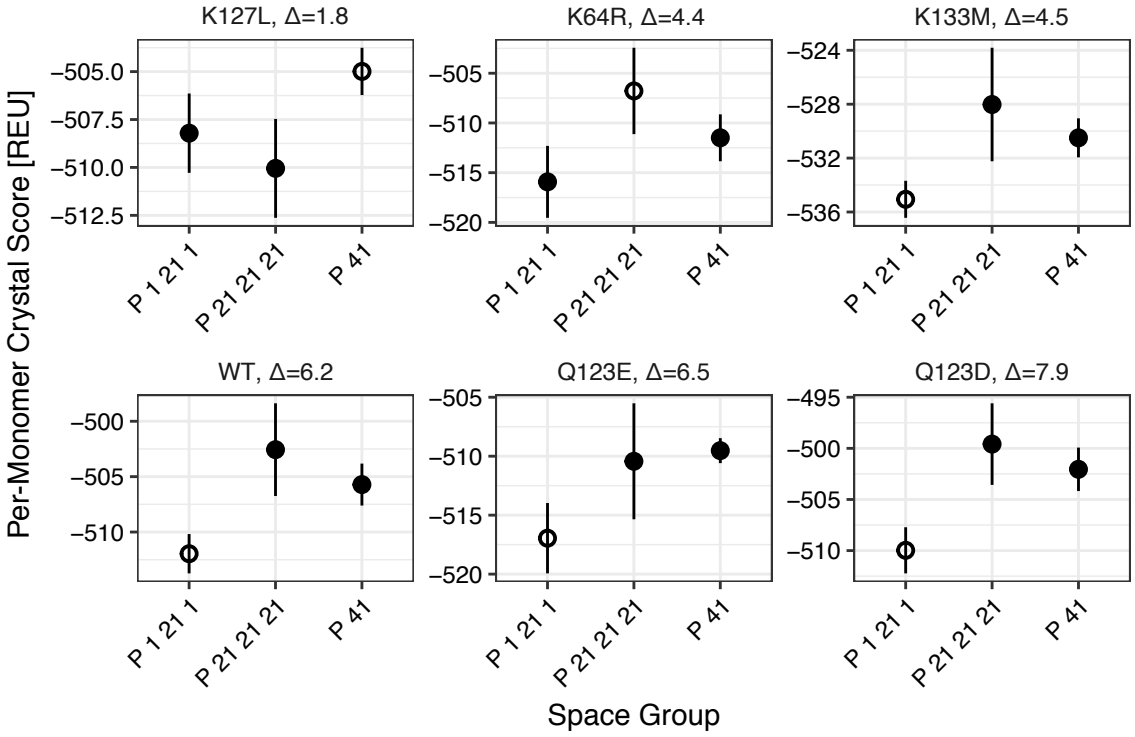

Figure 6.8: Scores of designs in the three most popular space groups for SNase. The lowest scoring space group is the experimentally observed space group in four out of six cases. For each design, average scores are shown $\pm$ one standard deviation, computed over ten energy-minimized structures in $P2_1$, $P2_12_12_1$, and $P4_1$. The experimentally observed space group is indicated by an open circle. The designs are ordered by the score difference, Δ, between the lowest scoring and second lowest scoring space group. The Δ value was greater on average for the designs crystallizing in $P2_1$ (6.3 REU vs. 3.6 REU). For, the two designs (K127L and K64R) that did not crystallize in $P2_1$, the correct space group could not be identified based on score.

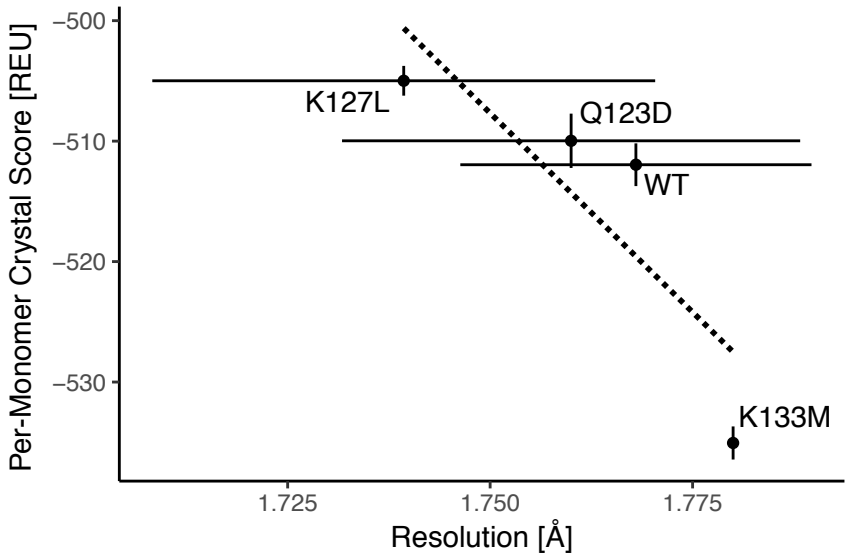

Figure 6.9: *Ex post facto* analysis shows an unexpected anti-correlation between resolution and score when the score is derived from the solved crystal structures of the designs. Error bars show the standard deviation in resolution (from collecting and analyzing multiple diffraction patterns) and score (from ten repeated energy-minimizations).

enhancing mutations at a rate of at least 11.5% (and at best 35.3%). I tested my design approach on a model SNase system, Δ+PHS, and found that my initial approach only resulted in one crystallizable variant out of ten, so I improved my approach by designing on an ensemble of backbones to increase this ratio to four in five. Finally, I solved the crystal structures of several of my designs but unfortunately observed little to no improvement in resolution. *Post-facto* analysis revealed that (1) improvements in resolution came primarily from changes in space group and (2) Rosetta score of designs was not predictive of crystal resolution.

6.5.1 Point mutations affected side-chain interactions and space groups

In general, when all variants were compared to both the WT and predicted design structures, the changes in the fold of SNase were undetectable, but there were detectable changes in side-chain interactions at the crystallographic interfaces. First, there were minimal changes in backbone structure, as anticipated for variants that differ by only point mutations at surface residues. The maximum observed root-mean-squared deviation (RMSD) for

backbone atoms (N, C, CA, O) between the designed model (or WT, since the backbone was fixed during design) and variant crystal structure was 0.53 Å for the variant K127L, with all other variants having lower backbone RMSD to their respective designed model (or WT, Supplemental Figure 6.A.2). Second, all mutations had some unpredicted effects on the interactions at the targeted crystallographic interface. These effects ranged from slight differences in side-chain rotameric states to entirely new interfaces. The smallest number of differences was observed in the K133M variant, where only a few side-chain dihedral angles differed from the designed structure and the interface was not greatly perturbed in general (Figure 6.5).

The greatest difference I observed was that two variants crystallized in higher symmetry space groups than WT ($P2_1$): K64R crystallized in the space group $P2_12_12_1$ and K127L crystallized in the space group $P4_1$. Several observed improvements in resolution were seen for this additional symmetry, such as a consistently higher I/σ value (a measure of the information content) over all resolution shells (Supplemental Figure 6.A.5). The space group change for K127L was driven by breaking the WT lysine–THP contact across one crystallographic interface (Figure 6.6), whereas the driver for the space group change of K64R was not definitively determined. The K64R substitution was desired because, in the WT space group (Figure 6.7), it introduced a putative electrostatic interaction between a terminal amino group of R64 and the carbonyl oxygen of K78 of a symmetry mate. The substitution did not disrupt any contacts, as K64 was not resolved in the electron density of the WT crystal structure (Supplemental Figure 6.A.4); however, this alteration still resulted in a change in space group. Since I was unable to solve a crystal structure for this variant, I resorted to modeling K64R in the new space group. My models hinted that this space group might be preferred over the WT because R64 can potentially form a hydrogen bond with E135 via both side-chain–side-chain and side-chain–backbone interactions (Figure 6.7).

6.5.2 The relationship between score and resolution is unclear

My initial hypothesis was that crystal interface stability, as captured by Rosetta score, would correlate with crystal resolution. I had two reasons that led me to this hypothesis. First, crystal growth occurs when the rate of protein incorporation into a crystal lattice (attachment) is greater than the rate of protein detachment from the crystal. The rate of attachment depends on the flux of molecules to the growing crystal step, the possible interaction area of the step, and the probability of attachment. The rate of detachment depends on the frequency of detachment events and the detachment probability. Point mutations can affect both the detachment and attachment probabilities. The detachment probability in particular is related to interface stability by the Boltzmann factor[42]: $P_d = e^{\frac{-E_i}{kT}}$. All other factors being equal, more stable native crystal interfaces will result in lower probabilities of detachment and thus may improve crystal morphology. Second, as interface stability is conferred by favorable molecular interactions and tighter side-chain packing, I reasoned that more stable interfaces would be less dynamic and less mobile, improving the homogeneity of protein positioning within the crystal. Hence, I anticipated that more stable interfaces would result in larger crystals with less mosaicity, which in turn would improve crystal resolution.

An initial analysis comparing Rosetta scoring and crystal resolution for a subset of the PDB revealed no relationship between the two. I reasoned that the analysis was obfuscated by many factors that affected resolution, but could not be captured by score alone (*e.g.* the protein, X-ray source intensity, detector resolution, user handling, *etc.*). First, I controlled for just the protein by analyzing only structures of SNase in the PDB. I found that controlling for the protein alone was insufficient – there was no trend between resolution and score for this set. However, when I controlled for more factors by analyzing the crystal structures of twenty-two variants of a single protein, with all data gathered by the same individual, using the same process, and with the same equipment, I found, as hypothesized, that lower Rosetta scores correlated with higher resolution (Figure 6.1). Yet, when I repeated the same

analysis for crystals of my model protein (again with all experiments conducted identically by one individual), I found an anti-correlation.

Why is there an inconsistent behavior between score and resolution? From the PDB, it is apparent that higher resolution crystal structures tend to have better protein geometry, *i.e.* fewer improbable side-chain rotamers, fewer outliers for bond lengths, fewer outliers for bond angles, or fewer atomic/steric clashes[43]. It is possible that because Rosetta represents proteins in internal coordinate space (Φ/Ψ), with fixed bond lengths and angles, it cannot rescue inherently poor geometry and thus better geometry contributes to a lower Rosetta score, even after energy minimization. Then, it is possible that the correlation observed between Rosetta score and resolution for the Mizutani set was a manifestation of the protein geometry improvements that come with higher resolution data, while some external factor, unaccounted for by Rosetta score, affected resolution. If resolution does indeed drive score, then for crystals in a narrow range of resolutions, we would not expect to observe a correlation between Rosetta score and resolution, as was the case for the Δ+PHS variants. In fact, MolProbity[44], a structure validation software, only compares structures within 0.25 Å bins to account for the improvements in protein geometry offered by higher resolution data.

6.5.3 Backrub improves design

Over the course of this study, I attempted both fixed-backbone design on the WT backbone and fixed-backbone design on a perturbed ensemble of 200 models generated from the WT backbone. To generate the perturbed ensemble, I used an approach known as Backrub[26] that slightly alters the direction of the Cα–Cβ vector to expose the side chain to a new environment while minimally altering the backbone. I found that variants designed using an ensemble of backbones crystallized at a higher rate (4/7) than variants design using the single WT backbone (1/11). I reason that this is because Backrub-generated ensembles capture local backbone fluctuations, whereas fixed-backbone models do not, resulting in

a better estimate of point mutation effects, including to interface energy[27]. In general, it is well known that proteins are dynamic and are readily capable of incorporating point mutations, especially at protein surface positions[45], so a fixed-backbone approximation is not sufficient. Hence, I observed an increase in crystallization (success) rate when I designed on an ensemble of backbones and selected designs scoring well across multiple backbones for experimental characterization.

6.5.4 Rosetta could not predict changes in rotamers and space groups

Of the five designed proteins, only K133M resulted in a crystal structure similar to the design. The designs Q123D and Q123E resulted in the introduction of E/D–K electrostatic interactions, but with a neighboring symmetry mate instead of the one targeted by the design. For these designs, it is not clear how to improve the design algorithm.

For the K127L and K64R designs, I observed unpredictable changes in space group. In retrospect, the K127L design should not have scored well in Rosetta, as the K127 amino group clearly makes electrostatic contacts with the phosphate groups of the THP molecule bound by the neighboring symmetry mate. Despite breaking the lysine–THP interaction, the Rosetta score was lower for the variant than the WT protein (in the WT space group), indicating that Rosetta does not correctly weigh the strength of this electrostatic interaction. One possible solution to overcome this issue in the future would be to bias the Rosetta score by the WT electron density, such that eliminating a clearly present interaction is strongly penalized whereas designing residues that are not well-resolved is favored.

One possible reason for the failure to improve resolution is that Rosetta is not yet finely tuned for the types of atomic interactions we tried to create. Rosetta was first developed to study protein folding in the context of small, globular domains, before being applied to the inverse challenge, design[46]. Over the years Rosetta has performed best when redesigning protein cores and tightly-packed interfaces[47–50]. Our objective here is one of the first attempts to design a loosely-packed interface with a significant amount of water involved.

Future work to improve design of solvated interfaces might include explicitly analyzing water interactions the interface either by flooding, as was recently successfully used to dock interfacial waters[51], or the recently developed HBNet method in Rosetta, which has been used to design hydrogen bonding networks *de novo*[52]. Multi-state design[53] might also be necessary to prevent undesired changes in space group.

6.5.5 The model protein was likely optimal for crystallization

Initial analyses showed that diffraction patterns collected for the WT control in this study had an average high-resolution limit of 1.77 Å (Figure 3), which agrees with the resolution (1.8 Å) of the PDB-deposited crystal structure of Δ+PHS (3BDC). This value falls firmly in the middle of the distribution of all SNase crystal resolutions (Figure 2), with 1.35 Å and 2.5 Å being the highest and lowest observed resolutions, respectively. Separately, Mizutani *et al.* observed changes from $+0.2$ to -0.6 Å in their study of the effects of point mutations on diphthine synthase crystals[17]. Based on these prior observations, I expected to observe changes in resolution of ± 0.5 Å; however, I instead found that designs spanned a narrow range of 1.67–1.85 Å or $\sim 1.77 \pm 0.1$ Å. Nonetheless, the variance in resolution within crystals of the same variant compared favorably between my study and that by Mizutani *et al.* I observed ranges of ~ 0.1 Å, while Mizutani *et al.* reported a 95% confidence interval estimate of ± 0.05 Å for WT diphthine synthase, analyzing the diffraction data from 13 crystals[17].

One possible explanation for my designs' minimal improvement in resolution is that my choice in model protein, Δ+PHS, was already optimized for forming high-quality crystals. I selected Δ+PHS as a model system for its high stability (11.8 kcal/mol)[29], high yield (over 60 mg protein per 1 L of cell culture), and crystallizability (over 300 crystal structures have been deposited in the PDB). I selected for these features so my model protein would readily incorporate point mutations and so the corresponding designs would likely express in high quantities and readily crystallize. However, these features also likely pre-selected for a protein that is optimal for crystallization, one for which a majority point mutations might

not be able to yield significant improvements in resolution. Future work might then focus on a protein that is less engineered and may have more room for improvement.

6.A Appendix

6.A.1 Supplemental figures

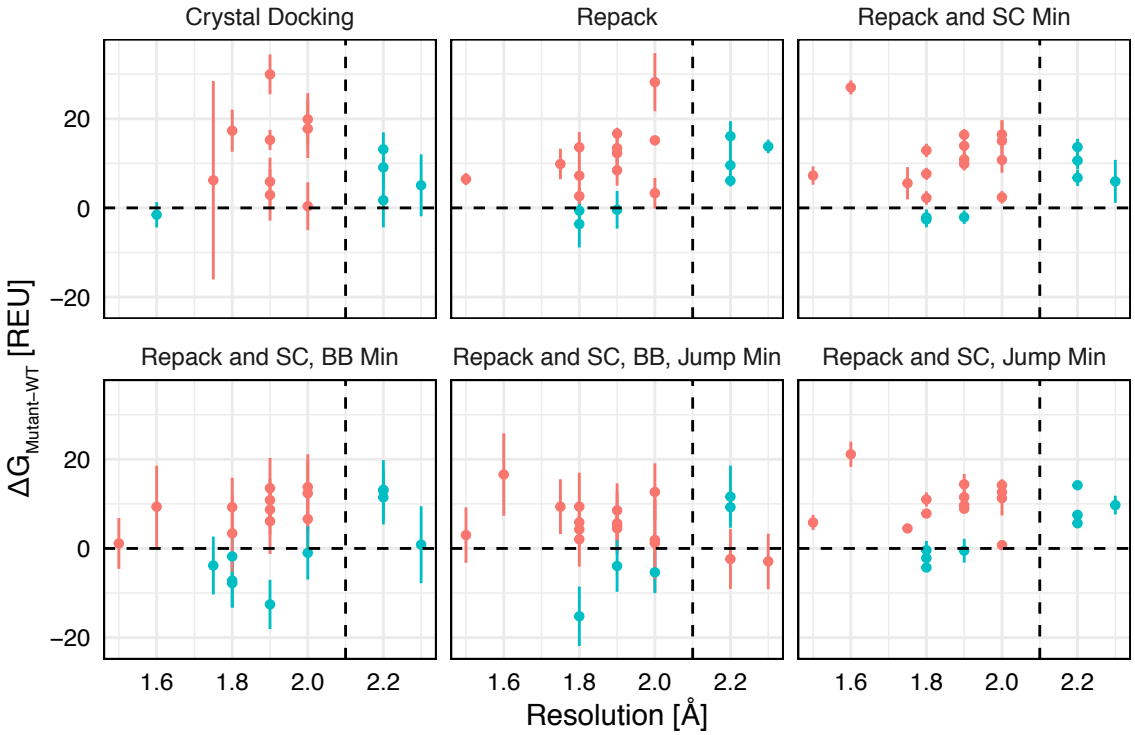

Figure 6.A.1: The ability for Rosetta to correctly forward design resolution-enhancing mutations on diphthine synthase depends on the degrees of freedom sampled during the simulation. Each plot here shows the same data as Figure 6.2 for a particular design strategy. The x-axis indicates the variant resolution and the y-axis shows the difference in energy with respect to the WT (hence the dashed lines represent the WT values for both). Standard deviations are calculated across 10 repeated design simulations. Point color indicates either a correct (*i.e.* a lower score than WT and corresponding higher resolution, blue) or incorrect prediction (red). The strategies are fully detailed in the Results section. Neither the strategy with the most (*Crystal Docking*, #6 in Results) nor the fewest (*Repack*, #1) degrees of freedom sampled was particularly successful. In fact, minimizing on side-chain (*Repack and SC Min*, #2) or rigid-body degrees (*Repack and SC, Jump Min*, #4) of freedom was not sufficient. Rather, successful strategies featured backbone minimization after the point mutation was introduced (*Repack and SC, BB Min*, #3, or *Repack and SC, BB, Jump Min*, #5).

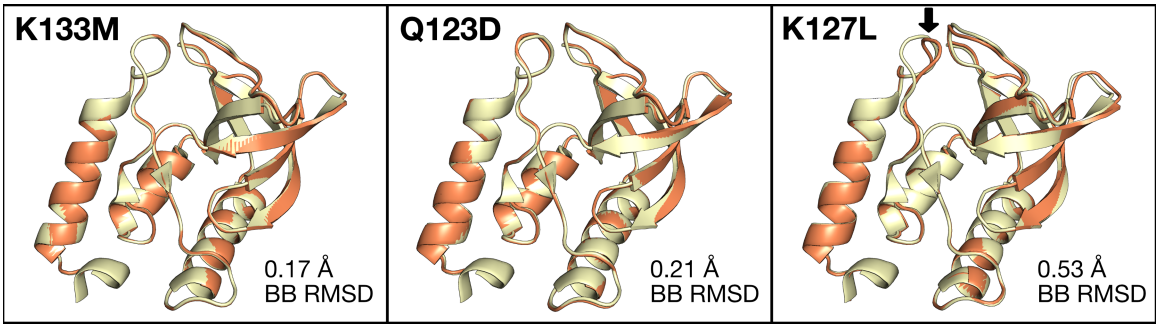

Figure 6.A.2: Alignments between designs (orange) and crystal structures (yellow) show minor backbone differences, with the greatest backbone RMSD being 0.53 Å for the K127L variant.

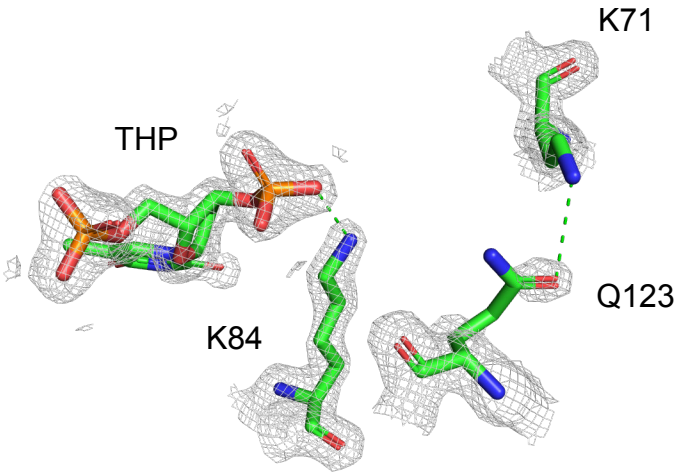

Figure 6.A.3: WT interactions between THP and K84, and Q123 and K71. Coordinates for the WT come from PDB ID 3BDC. The 2mFo-DFc map was downloaded from the Uppsala Electron Density Server. The map is contoured at 1.5σ and carved within 2 Å of residues 71, 84, 123, and the THP. Distances below 3.5 Å are highlighted by green dashed lines.

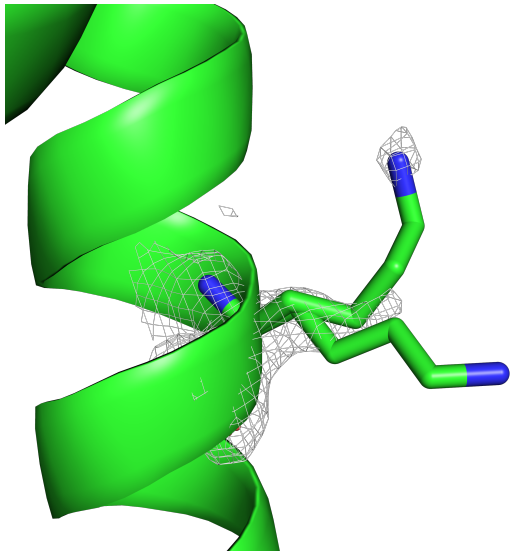

Figure 6.A.4: WT density of K64 is missing for some atoms and shows multiple rotameric states. Coordinates for the WT come from PDB ID 3BDC. The 2mFo-DFc map was downloaded from the Uppsala Electron Density Server. The map is contoured at 1.5σ and carved within 2 Å of residue 64.

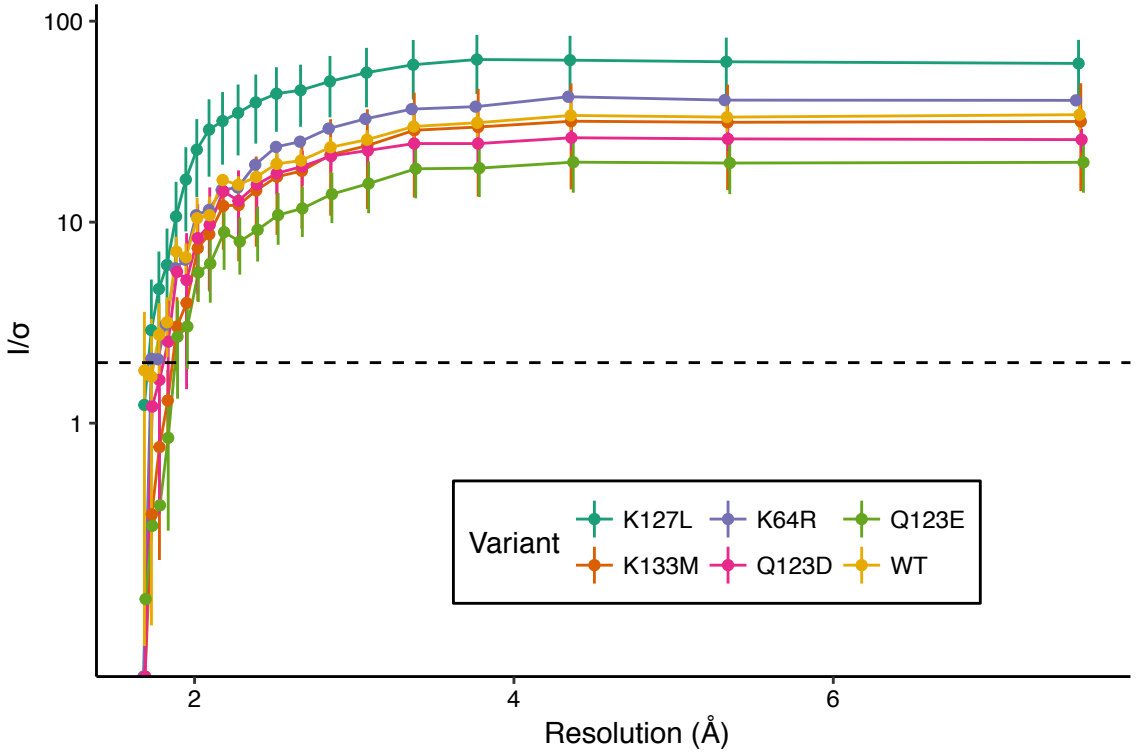

Figure 6.A.5: Average I/σ values $\pm$ one standard deviation for each resolution shell for each variant. Averages and standard deviations are calculated over all protein crystals for which diffraction data could be collected.

6.A.2 Supplemental tables

Table 6.A.1: Exact crystallization conditions for each crystal from which diffraction data were collected, including initial values for I/σ and completeness in the last resolution shell

Variant	% MPD	pH	Ca2+ Ratio	pdTp Ratio	Space Group	Resolution	I/σ	Completeness (%)
Δ+PHS K127L	40	9	3	2	P 41	1.67	3.36	9.0
Δ+PHS K127L	40	9	3	2	P 41	1.77	4.74	48.8
Δ+PHS K127L	40	8	3	2	P 41	1.69	2.37	9.5
Δ+PHS K127L	42	9	3	2	P 41	1.77	2.23	46.3
Δ+PHS K127L	42	9	3	2	P 41	1.69	5.10	9.0
Δ+PHS K127L	42	9	3	2	P 41	1.68	3.59	7.7
Δ+PHS K127L	42	8	3	2	P 41	1.73	4.78	25.0
Δ+PHS K127L	44	9	3	2	P 41	1.83	2.72	79.6
Δ+PHS K127L	44	8	3	2	P 41	1.73	4.57	23.2
Δ+PHS K127L	44	8	3	2	P 41	1.78	2.48	44.0
Δ+PHS K127L	46	9	3	2	P 41	1.68	2.98	7.7
Δ+PHS K127L	46	9	3	2	P 41	1.73	6.77	26.4
Δ+PHS K127L	46	8	3	2	P 41	1.68	2.29	8.7
Δ+PHS K127L	46	8	3	2	P 41	1.83	3.05	69.5
Δ+PHS K127L	46	8	3	2	P 41	1.78	2.34	46.4
Δ+PHS WT	21	6	3	2	P 1 21 1	1.73	3.35	11.4
Δ+PHS WT	18	6	3	2	P 1 21 1	1.77	3.33	33.0
Δ+PHS WT	18	6	3	2	P 1 21 1	1.78	2.74	25.9
Δ+PHS WT	18	6	3	2	P 1 21 1	1.74	3.60	12.1
Δ+PHS WT	18	6	3	2	P 1 21 1	1.83	2.08	45.4
Δ+PHS Q123E	18	6	2	1	P 1 21 1	1.89	2.79	68.8
Δ+PHS Q123E	20	6	2	1	P 1 21 1	1.94	2.62	66.2
Δ+PHS Q123E	20	6	2	1	P 1 21 1	1.89	3.61	56.5
Δ+PHS Q123E	20	6	2	1	P 1 21 1	1.84	2.37	24.6
Δ+PHS Q123E	18	6	2	1	P 1 21 1	1.89	5.14	60.2
Δ+PHS Q123E	18	6	2	1	P 1 21 1	1.88	2.64	67
Δ+PHS Q123E	18	6	2	1	P 1 21 1	1.96	2.37	87.1
Δ+PHS K133M	20	6	2	1	P 1 21 1	1.84	2.16	43.7
Δ+PHS K133M	18	6	2	1	P 1 21 1	1.89	2.48	67.4
Δ+PHS K133M	18	6	2	1	P 1 21 1	1.90	3.98	81.6
Δ+PHS K133M	18	6	2	1	P 1 21 1	1.79	2.62	26.6
Δ+PHS K133M	22	6	2	1	P 1 21 1	1.89	3.58	74.3
Δ+PHS Q123D	20	6	2	1	P 1 21 1	1.95	2.39	87.9
Δ+PHS Q123D	20	6	2	1	P 1 21 1	1.78	2.61	25.9
Δ+PHS K64R	18	6	3	2	P 21 21 21	1.73	2.13	22.2

Table 6.A.2: Crystallographic data collection and refinement statistics. Values for the highest resolution shell are in parenthesis. PDB IDs to be added after submission.

	K127L (6OK8)	K133M (6OK9)	Q123D (6OKA)
Wavelength (Å)	1.54	1.54	1.54
Resolution Range (Å)	38.26–1.80 (1.87–1.80)	32.21–1.90 (1.97–1.90)	32.3–1.86 (1.93–1.86)
Space Group	$P4_1$	$P2_1$	$P2_1$
Unit Cell Dimensions			
a, b, c (Å)	48.097 48.097 63.122	30.927 60.473 38.105	30.853 60.715 38.119
α, β, γ **(°)**	90 90 90	90 93.017 90	90 92.648 90
Total Reflections	72360 (1421)	33219 (1633)	34560 (1407)
Unique Reflections	12307 (715)	10702 (874)	11133 (680)
Multiplicity	5.9 (2.0)	3.1 (1.9)	3.1 (2.1)
Completeness (%)	91.93 (53.32)	96.32 (79.71)	93.59 (56.06)
Mean I/σ(I)	35.73 (5.59)	25.62 (3.92)	18.37 (5.10)
Wilson B-factor	22.28	29.02	28.23
R-merge	0.02957 (0.1052)	0.02397 (0.1765)	0.03833 (0.1442)
R-meas	0.03219 (0.1381)	0.02873 (0.2304)	0.04588 (0.1879)
R-pim	0.01254 (0.08851)	0.01565 (0.146)	0.02494 (0.119)
CC1/2	1 (0.984)	0.999 (0.983)	0.998 (0.978)
CC*	1 (0.996)	1 (0.996)	1 (0.994)
Reflections used			
in refinement	12308 (715)	10706 (868)	11125 (680)
Reflection used			
for R-free	619 (38)	538 (44)	562 (38)
R-work	0.1950 (0.2892)	0.2199 (0.3940)	0.1986 (0.3391)
R-free	0.2331 (0.4364)	0.2635 (0.4842)	0.2532 (0.4043)
CC(work)	0.961 (0.826)	0.956 (0.785)	0.935 (0.475)
CC(free)	0.947 (0.871)	0.953 (0.818)	0.941 (0.205)
# non-H Atoms:			
Total	1147	1058	1102
Macromolecules	1032	993	999
Ligands	26	26	26
Solvent	89	39	77
Protein Residues	129	129	129
RMS(bonds) Å	0.022	0.025	0.022
RMS(angles) °	2.03	1.63	1.73
Ramachandran			
favored (%)	92.91	94.49	93.70
allowed (%)	4.72	4.72	5.51
outliers (%)	2.36	0.79	0.79
Rotamer outliers (%)	3.74	3.09	1.04
Clashscore	1.41	2.52	3.01
Average B-factor:			
Total	25.95	38.52	31.71
Macromolecules	25.29	38.54	31.51
Ligands	34.24	39.44	28.08
Solvent	31.23	37.34	35.60

References

1. Giegé, R. A historical perspective on protein crystallization from 1840 to the present day. *The FEBS journal* **280,** 6456–97 (2013).
2. Fusco, D. *et al.* Statistical analysis of crystallization database links protein physicochemical features with crystallization mechanisms. *PLoS ONE* **9,** e101123 (2014).
3. Price II, W. N. *et al.* Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. en. *Nature Biotechnology* **27,** 51–57 (2009).
4. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242 (2000).
5. Richardson, J. S. *The anatomy and taxonomy of protein structure* (Academic Press, 1981).
6. Knight, S, Andersson, I & Brändén, C. I. Reexamination of the Three-Dimensional Structure of the Small Subunit of RuBisCo from Higher Plants. en. *Science (New York, N.Y.)* **244,** 702–5 (1989).
7. Rupp, B & Segelke, B. Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide. *Nature structural biology* **8,** 663–4 (2001).
8. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS journal* **275,** 1–21 (2008).
9. Dale, G. E., Oefner, C. & D'Arcy, A. The protein as a variable in protein crystallization. *Journal of Structural Biology* **142,** 88–97 (2003).
10. Evdokimov, A. G. *et al.* Rational protein engineering in action: The first crystal structure of a phenylalanine tRNA synthetase from Staphylococcus haemolyticus. *Journal of Structural Biology* **162,** 152–169 (2008).
11. Lawson, D. M., Smith, J. M. A. & et al. Solving the Structure of Human H Ferritin by Genetically Engineering Intermolecular Crystal Contacts. en. *Nature* **349,** 541–544 (1991).
12. Cooper, D. R. *et al.* Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta crystallographica. Section D, Biological crystallography* **63,** 636–45 (2007).
13. Lanci, C. J. *et al.* Computational design of a protein crystal. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 7304–7309 (2012).

14. Oubridge, C., Ito, N., Teo, C.-H. H., Fearnley, I. & Nagai, K. Crystallisation of RNA-protein complexes. II. The application of protein engineering for crystallisation of the U1A protein-RNA complex. *Journal of molecular biology* **249,** 409–423 (1995).

15. Lai, Y.-T. *et al.* Lattice engineering enables definition of molecular features allowing for potent small-molecule inhibition of HIV-1 entry. *Nature Communications* **10,** 47 (2019).

16. Fusco, D. *et al.* Characterizing protein crystal contacts and their role in crystallization: rubredoxin as a case study. en. *Soft Matter* **10,** 290–302 (2013).

17. Mizutani, H. *et al.* Systematic study on crystal-contact engineering of diphthine synthase: Influence of mutations at crystal-packing regions on X-ray diffraction quality. *Acta Crystallographica Section D: Biological Crystallography* **64,** 1020–1033 (2008).

18. Wang, G. & Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics* **19,** 1589–1591 (2003).

19. Bliven, S., Lafita, A., Parker, A., Capitani, G. & Duarte, J. M. Automated evaluation of quaternary structures from protein crystals. *PLOS Computational Biology* **14** (ed Dunbrack, R. L.) e1006104 (2018).

20. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science* **23,** 47–55 (2014).

21. Nivon, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS ONE* **8** (ed Zhang, Y.) e59004 (2013).

22. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling Symmetric Macromolecular Structures in Rosetta3. *PLoS ONE* **6** (ed Uversky, V. N.) e20450 (2011).

23. Chaudhury, S., Lyskov, S. & Gray, J. J. *PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta* 2010.

24. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13,** 3031–3048 (2017).

25. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation* **12,** 6201–6212 (2016).

26. Smith, C. A. & Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *Journal of Molecular Biology* **380,** 742–756 (2008).

27. Barlow, K. A. *et al.* Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *The Journal of Physical Chemistry B* **122,** 5389–5399 (2018).

28. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnology* **8,** 91 (2008).

29. Castañeda, C. A. *et al.* Molecular determinants of the pKavalues of Asp and Glu residues in staphylococcal nuclease. *Proteins: Structure, Function and Bioinformatics* **77,** 570–588 (2009).

30. García-Moreno, B. E. *et al.* Experimental measurement of the effective dielectric in the hydrophobic core of a protein. *Biophysical Chemistry* **64,** 211–224 (1997).

31. Kabsch, W. XDS. *Acta Crystallographica Section D Biological Crystallography* **66,** 125–132 (2010).

32. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography* **66,** 213–221 (2010).

33. McCoy, A. J. *et al.* Phaser crystallographic software. *Journal of Applied Crystallography* **40,** 658–674 (2007).

34. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography* **66,** 486–501 (2010).

35. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D: Biological Crystallography* **68,** 352–367 (2012).

36. DiMaio, F. *et al.* Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **473,** 540–543 (2011).

37. Leaver-Fay, A., Snoeyink, J. & Kuhlman, B. *On-the-fly rotamer pair energy evaluation in protein design* in *Lecture Notes in Computer Science* **4983** (Berlin, Heidelberg, 2008), 343–354.

38. Benjamin Stranges, P. & Kuhlman, B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Science* **22,** 74–82 (2013).

39. Evans, P. & IUCr. Scaling and assessment of data quality. *Acta Crystallographica Section D Biological Crystallography* **62,** 72–82 (2006).

40. Karplus, P. A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Current Opinion in Structural Biology* **34,** 60–68 (2015).

41. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).

42. De Yoreo, J. J. Principles of Crystal Nucleation and Growth. *Reviews in Mineralogy and Geochemistry* **54,** 57–93 (2003).

43. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27,** 293–315 (2018).

44. Chen, V. B. *et al.* MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66,** 12–21 (2010).

45. Davis, I. W., Arendall, W. B., Richardson, D. C. & Richardson, J. S. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. *Structure* **14,** 265–274 (2006).

46. Baker, D. What has de novo protein design taught us about protein folding and biophysics? *Protein Science* **28,** 678–683 (2019).

47. Kuhlman, B. *et al.* Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. en. *Science* **302,** 1364–1368 (2003).

48. Kortemme, T. *et al.* Computational redesign of protein-protein interaction specificity. *Nature structural & molecular biology* **11,** 371–379 (2004).

49. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science (New York, N.Y.)* **357,** 168–175 (2017).

50. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science (New York, N.Y.)* **353,** 389–94 (2016).

51. Kilambi, K. P. *et al.* Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20-27. en. *Proteins: Structure, Function and Bioinformatics* **81,** 2201–2209 (2013).

52. Boyken, S. E. *et al.* De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science (New York, N.Y.)* **352,** 680–7 (2016).

53. Leaver-Fay, A., Jacak, R., Stranges, P. B. & Kuhlman, B. A generic program for multistate protein design. *PloS one* **6,** e20937 (2011).

Chapter 7

Conclusion and Future Directions

Proteins are the machinery of life, involved in most, if not all, cellular function. Structural studies of proteins provide snapshot images of these machines at work and are necessary to build a bottom-up understanding of biology. For example, acquiring atomic-resolution structural information about proteins involved in disease, such as antibodies, can lead to the development of new therapeutics and vaccines. In most cases, X-ray crystallography is the standard approach to structure determination, but it is not always feasible. X-ray crystallography may not be the most optimal approach when there are numerous protein targets because of its significant labor and reagent cost or when the target proteins are flexible and do not possess a single, static structure because of the requirement for structural homogeneity underlying the technique. Computational approaches that provide atomistic models can complement X-ray crystallography in these cases, as computational methods methods are inexpensive and often output a range of plausible models. In this dissertation, I have advanced computational modeling approaches for antibodies and antibody–antigen complexes, applied modeling to gain scientific insight to systems which would otherwise have been unattainable through experiment, and attempted to combine computation and experiment to improve the resolution of crystal structures.

7.1 My Contributions

I began my research in computational protein modeling through participation in the CAPRI competition[1]. CAPRI's numerous blind protein complex prediction challenges highlighted shortcomings in Rosetta's modeling and docking of camelid (heavy-chain only) antibodies. To address these shortcomings, I developed a more robust antibody modeling framework free from previous assumptions that were based on the traditional antibody structure of a paired heavy and light chain. Simultaneously, a growth in the need for antibody modeling exposed a lack of breadth in the template database and an unsustainability in our template grafting script. I addressed the former issue by developing an automatically updating database and implementing a scientific benchmark to evaluate grafting accuracy. The latter issue was addressed by refactoring the grafting protocol to be object-oriented, providing a framework for future development, by a team of developers, myself included[2].

My development of RosettaAntibody was inspired in part by recent improvements in the accuracy and decreases in the cost of high-throughput B cell[3] sequencing. As the recent growth in antibody sequences has not been matched by a growth in antibody structures, there was an opportunity to test whether modeling could be used to gain structural insights from a large set of antibodies. Experimentally-derived antibody sequences can be categorized as naïve or antigen experienced (based on cell-surface receptors). This distinction permitted me to ask whether the process of affinity maturation (antigen exposure) drives CDR-H3 loop rigidification, a structural property that had previously only been studied on the scale of tens of antibodies[4–6]. Coupling Rosetta modeling with a graph theoretical approach for quantifying flexibility[7] from a static structure, I determined the flexibility of the CDR-H3 loop for thousands of models of the human peripheral blood cell antibody repertoire. I found no clear delineation in the flexibility of naïve and antigen-experienced antibodies, contrary to prior observations. I further investigated this surprising result by using additional measures of flexibility and studying the hundreds of crystal structures

available in the PDB. Again, I did not observe a drastic decrease in flexibility upon affinity maturation. Further analysis still incorporated molecular dynamics and showed that there was a spectrum of changes in flexibility that depended on the specific antibody in question. My results suggested that rigidification may be just one of many biophysical mechanisms for increasing affinity, and were recently validated by another research group[8].

Another, similar[i], system I computationally modeled was the bacterial protein Hfq, which is present in most sequenced bacteria. With strong sequence conservation only in its core domain, the role of Hfq's termini is unclear. To investigate, I modeled *E. coli* Hfq and identified key interactions between its disordered C-terminal domains (CTDs) and its ordered core domains (Hfq is a homohexamer). In the process, I improved the disordered region modeling protocol in Rosetta, FloppyTail, by enabling the simultaneous modeling of multiple disordered regions, examining the extent of sampling in ultra-long simulations, and developing a novel, ensemble-based analysis for low-scoring models. For *E. coli* Hfq, I identified multiple key CTD–core interactions, which were validated experimentally. In conjunction with competitive binding experiments, the models showed that the acidic CTD transiently bound the basic core residues at the Hfq rim, which are involved in RNA annealing. To test whether this result was generalizable, I modeled the Hfq proteins found in five other bacterial species and showed that the presence of CTD–rim interactions was correlated with RNA annealing activity in four of the five species. Separately, my Hfq models for one of the species, *C. crescentus*, were validated by a recently determined crystal structure[9].

The final thrust of my PhD research focused on applying computational design to improve the resolution of protein crystal structures. I demonstrated that the resolution of a subset of crystal structures in the PDB correlated with the Rosetta score of the crystallographic interactions. In a "forward design" study, I investigated multiple computational design approaches on this subset of crystal structures to identify the one with the greatest

[i] In the sense that many variants of a single protein exist that would be tedious to study experimentally, but can be feasibly modeled.

success rate. I then applied this approach to design the crystallographic interactions of a model protein (SNase). Only five of my sixteen designs formed crystals, and of the crystal-forming designs only two slightly improved resolution. Surprisingly, after solving the designs' crystal structures, I found that two had altered space groups, which could not be predicted by Rosetta score, and that, over the narrow resolution range of the designs, score no longer correlated with resolution. My results show that point mutations can have significant effects on protein crystallization, but may have been hampered by my efforts to design a protein that already optimal for crystallization.

7.2 Future Directions

I will present future directions in reverse order from how the thesis is structured, starting with crystal design. I believe there is much potential for computational design to stabilize weak crystallographic interfaces. Chapter 6 showed that point mutations can have significant effects. However, design attempts only resulted in minimal improvements (∼0.1 Å) to resolution as the model protein crystallized quite readily, forming crystals that diffracted to a high resolution (∼1.8 Å). Going forward, I propose the computational redesign of ribonuclease H, a protein that is easy to purify in high quantities, but that does not form crystals diffracting beyond 2.8 Å[10]. Optimizing a protein crystal with an initial resolution of 2.8 Å may leave more room for improvement than a protein crystal starting at 1.8 Å. Future designs should consider making use of a large ensemble of backbones, as I demonstrated that using a Backrub-generated[11] ensemble improved design success rate. Another advance to be made beyond my original approach is to design the entire interface, which can take the form of fixed backbone design or include the addition of contact-forming loops, rather than targeting single point mutations.

Modeling disordered regions continues to be challenging, although I have shown in Chapter 5 that models can be predictive and, when combined with experiments, can elucidate biological mechanisms. A weakness of Rosetta FloppyTail is its lack of detailed balance,

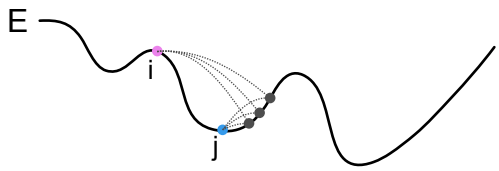

Figure 7.1: Schematic example of the effects of minimization on moves from state i, in energy landscape E. Minimization inevitable leads to state j, if a move is made to any intermediate state (gray, unlabelled).

which prevents the calculation of true thermodynamic properties from simulations. Detailed balance requires equal-probability sampling of states and is defined as $P_i\pi_{ij} = P_j\pi_{ji}$[12], where P_i is the probability of observing state i and π_{ij} is the probability of transitioning to state j from state i, which can also be written as $\pi_{ij} = \alpha_{ij}a_{ij}$, where α_{ij} is probability of performing the move from state i to state j and a_{ij} is the probability of accepting them move. FloppyTail, and most Rosetta protocols, break this by minimizing before evaluating the Metropolis criterion so π_{ij} and π_{ji} are skewed[ii]. See for example, Figure 7.1, where if a Rosetta move from state i goes to any state near the local energy minima j, minimization will move towards j, and this probability of sampling j is much greater than it would be under a sampling approach without minimization. Working to eliminate such bias in FloppyTail, a future developer would begin by creating a move set without minimization. Additionally, they would need to demonstrate fair sampling of conformational space by their move set as Smith and Kortemme have for the Backrub protocol[13]. A thermodynamically rigorous FloppyTail would permit more direct comparison to experiment and could be used to make stronger predictions (*e.g.* the occupancy of low-energy states would be known rather than just knowing that the states are low energy).

Further advances could be made in the physical rigor of the low-resolution energy potential of FloppyTail (and Rosetta in general). In Rosetta, the `cen_std` low-resolution

[ii] That's not to say this is the only way to break detail balance. Biasing sampling, for example, would also suffice as it effectively increases the probability transitioning to or occupying certain states.

energy potential contains only four terms: `env`, `pair`, `cbeta`, `vdw`[14]. The three first terms are purely statistical and capture all residue–residue interactions. The latter term is physical, but only contains the repulsive component of the Lennard-Jones potential. It is somewhat surprising that an energy potential with no explicit consideration of electrostatics has been successful in modeling disordered interactions, which are heavily dependent on electrostatics[15]. A step forward might be to incorporate additional physical terms in the low-resolution potential, while validating against experimental observations. An approach in this spirit has been made by John Ferrie (Department of Chemistry, University of Pennsylvania), who constrained disordered regions during simulation using a potential derived from the Gaussian chain probability distribution[16].

In Chapter 4 I showed how antibody modeling could be combined with antibody sequencing to garner useful structural information, despite the fact that modeling is not 100% accurate. The need for fast and accurate antibody modeling will rise over the coming years due to the development of high-throughput B-cell sequencing technologies[3,17]. Most antibody regions are already modeled at close to 90% accuracy and with reasonable speed[18,19]. The only exception is the CDR-H3 loop.

Modeling the CDR-H3 loop is slow because models must be generated *de novo* and refined. Most of the time in a loop closure simulation is not spent identifying plausible loop backbone conformations, but rather placing the residue side chains in a low-energy conformation for their environment and ruling out backbone conformations that result in side-chain clashes. The difference is an order of magnitude, with a 12-residue loop requiring approximately 200 seconds for *de novo* closure and 4,000 seconds for refinement. Thus to decrease the computational time cost of CDR-H3 loop modeling, one must expedite the high-resolution refinement stage of Rosetta's loop modeling protocol, or incorporate a different, fast loop modeling approach such as DiSGro[20,21], which is efficient because it samples only relevant protein conformations by building loops sequentially where each residues placement is based on observed distance distributions from the PDB.

In addition to improving the speed of CDR-H3 loop modeling, we must also strive to improve the accuracy, which is a significant challenge requiring two sources of error to be addressed. First, during antibody modeling, multiple templates are grafted into one structure and this introduces model error; the local environment for the loop does approximate the native environment well. Second, even when in a close-to-native environment, we struggle to model CDR-H3 loops, because these loops are inherently structurally diverse[22]. Overcoming the first source of error amounts to improving the template prediction. This can be done by testing template selection strategies that are more sensitive to minor sequence changes than BLAST (*e.g.* decision trees[19] or position-specific scoring matrices [PSSMs][23]). Implementing and benchmarking new grafting strategies should be straightforward in the current RosettaAntibody framework thanks to the advances I outlined in Chapter 3. Overcoming errors in CDR-H3 loop modeling will be a more difficult proposition. Perhaps a rigorous characterization of particularly challenging loops might yield some insight as to why these cases are so difficult and could lead to an improved modeling strategy. For example a long molecular dynamics simulation might reveal that the crystallographic loop conformation is but one of many accessible low-energy states, so modeling the loop as an ensemble would be an improved approach, particularly if one seeks to use the models for downstream applications such as docking.

Antibody–antigen docking itself could be improved. One exciting advance is being pursued, in Prof. Jeff Gray's lab, by Dr. Jing Zhou. She is currently developing a SnugDock variant that takes advantage of hydrogen–deuterium exchange mass spectrometry (HDX-MS) data to more accurately model camelid antibodies (cAb) and cAb–antigen complexes. It would be a substantial advance if combining SnugDock and HDX-MS could yield atomic accuracy models, because HDX-MS data is easier and faster to collect for a given CAb–antigen complex than it is to solve the crystal structure. Preliminary results show that experimentally-guided SnugDock simulations sample more low-scoring, native-like states, albeit with a corresponding increase in non-native low-scoring states. More work is

necessary to distinguish the two sets of states in a blind scenario. However, the concept of combining experimental data with simulation to improve accuracy is a powerful one, and more efforts should be made to standardize this incorporation as high-throughput experiments are steadily becoming the norm and data will be plentiful.

7.3 Parting Thoughts

To unite multiple, disparate research topics in a single dissertation speaks volumes on the current state of computation and science in general. Rosetta's versatility has grown exponentially since its inception twenty-two years ago[24] as a protein folding tool. Nowadays, it is possible to model not just soluble, globular proteins, but also non-canonical amino acids[25], RNA[26], membrane proteins[27], and carbohydrates[28]. These advances have been enabled by the conversion of Rosetta from a set of Fortran subroutines to an easy-to-use and well-organized object-oriented C++ framework[29] and by the development of a more rigorous energy function[30]. The versatility of Rosetta will only continue to grow as Python[31] and XML[32] scripting interfaces expose the underlying code to regular users. The universal accessibility of modeling tools will be key as scientific research is becoming ever more interdisciplinary. If one thing has been demonstrated in this dissertation it is that by combining computational and experimental approaches, we are able to solve more diverse set of problems than when using either approach in isolation.

References

1. Marze, N. A. *et al.* Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28-35. *Proteins: Structure, Function, and Bioinformatics* **85,** 479–486 (2017).
2. Weitzner, B. D. *et al.* Modeling and docking of antibody structures with Rosetta. *Nature Protocols* **12,** 401–416 (2017).
3. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. en. *Nature Biotechnology* **32,** 158–168 (2014).
4. Jimenez, R., Salazar, G., Baldridge, K. K. & Romesberg, F. E. Flexibility and molecular recognition in the immune system. *Proc Natl Acad Sci U S A* **100,** 92–97 (2003).
5. Thorpe, I. F., Brooks, C. L. & Brooks 3rd, C. L. Molecular evolution of affinity and flexibility in the immune system. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 8821–8826 (2007).
6. Willis, J. R. *et al.* Human germline antibody gene segments encode polyspecific antibodies. *PLoS Comput Biol* **9,** e1003045 (2013).
7. Sljoka, A. *Algorithms in rigidity theory with applications to protein flexibility and mechanical linkages* PhD thesis (York University, 2012).
8. Ovchinnikov, V., Louveau, J. E., Barton, J. P., Karplus, M. & Chakraborty, A. K. Role of framework mutations and antibody flexibility in the evolution of broadly neutralizing antibodies. *eLife* **7** (2018).
9. Santiago-Frangos, A. *et al.* Caulobacter crescentus Hfq structure reveals a conserved mechanism of RNA annealing regulation. *Proceedings of the National Academy of Sciences of the United States of America,* 201814428 (2019).
10. Ishikawa, K. *et al.* Crystal structure of ribonuclease H from Thermus thermophilus HB8 refined at 2.8 A resolution. *Journal of Molecular Biology* **230,** 529–542 (1993).
11. Barlow, K. A. *et al.* Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *The Journal of Physical Chemistry B* **122,** 5389–5399 (2018).
12. Frenkel, D. & Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications* (2002).
13. Smith, C. A. & Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *Journal of Molecular Biology* **380,** 742–756 (2008).

14. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods in enzymology* **383,** 66–93 (2004).

15. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences* **110,** 13392–13397 (2013).

16. Ferrie, J. J. *et al.* Using a FRET Library with Multiple Probe Pairs To Drive Monte Carlo Simulations of alpha-Synuclein. *Biophysical Journal* **114,** 53–64 (2018).

17. Kovaltsuk, A. *et al.* How B-cell receptor repertoire sequencing can be enriched with structural antibody data. *Frontiers in Immunology* **8,** 1753 (2017).

18. Weitzner, B. D., Dunbrack, R. L. & Gray, J. J. The origin of CDR H3 structural diversity. *Structure* **23,** 302–11 (2015).

19. Long, X., Jeliazkov, J. R. & Gray, J. J. Non-H3 CDR template selection in antibody modeling through machine learning (2018).

20. Tang, K., Zhang, J. & Liang, J. Fast Protein Loop Sampling and Structure Prediction Using Distance-Guided Sequential Chain-Growth Monte Carlo Method. *PLoS Computational Biology* **10,** e1003539 (2014).

21. Tang, K., Zhang, J. & Liang, J. Distance-guided forward and backward chain-growth Monte Carlo method for conformational sampling and structural prediction of antibody CDR-H3 loops. *Journal of Chemical Theory and Computation* **13,** 380–388 (2017).

22. Regep, C., Georges, G., Shi, J., Popovic, B. & Deane, C. M. The H3 loop of antibodies shows unique structural characteristics. *Proteins: Structure, Function and Bioinformatics* **85,** 1311–1318 (2017).

23. Wong, W. K. *et al.* SCALOP: sequence-based antibody canonical loop structure annotation. *Bioinformatics* (ed Valencia, A.) (2018).

24. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology* **268,** 209–225 (1997).

25. Renfrew, P. D., Choi, E. J., Bonneau, R. & Kuhlman, B. Incorporation of Noncanonical Amino Acids into Rosetta and Use in Computational Protein-Peptide Interface Design. *PLoS ONE* **7** (ed Uversky, V. N.) e32637 (2012).

26. Das, R. Atomic-Accuracy Prediction of Protein Loop Structures through an RNA-Inspired Ansatz. *PLoS ONE* **8,** e74830 (2013).

27. Alford, R. F. *et al.* An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLOS Computational Biology* **11** (ed Livesay, D. R.) e1004398 (2015).

28. Labonte, J. W., Adolf-Bryfogle, J., Schief, W. R. & Gray, J. J. Residue-centric modeling and design of saccharide and glycoconjugate structures. *Journal of Computational Chemistry* **38,** 276–287 (2017).

29. Leaver-Fay, A. *et al.* Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology. Computer Methods, Part C* **487** (eds Brand, M. L. J. & Ludwig) 545–574 (2011).

30. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation* **12,** 6201–6212 (2016).

31. Chaudhury, S., Lyskov, S. & Gray, J. J. *PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta* 2010.

32. Fleishman, S. J. *et al.* RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLoS ONE* **6** (ed Uversky, V. N.) e20161 (2011).

Vita

Jeliazko R. Jeliazkov was born in a town in the Eastern Rhodopes on a rather cold day in 1992. He wishes he could be more specific in this description, but fears that if he were, the secret questions to any one of his bank or email accounts would not longer be secret. Growing up, he found that his parents, a doctor and an engineer, worked too hard for his liking, so he decided to pursue an academic career.

He had his first, rather harrowing research experience at the age of seventeen with Prof. Julia Velkovska (Department of Physics and Astronomy, Vanderbilt University). As it turned out, writing C++ code to analyze heavy ion collisions was exceptionally challenging for a novice physics student. Reconsidering his options (and calculating the probability that his parents would fund a degree in Comparative Literature), he decided to give physics another try, but after he had taken some more classes. In his final undergraduate years, he worked with Prof. Karin Dahmen, studying slip avalanche statistics in granular systems, and Eugene Colla, studying phase transition kinetics in ferroelectric relaxors, both in the Department of Physics at the University of Illinois at Urbana-Champaign. This time his studies paid off and he was rewarded with a second-author publication.

His scientific appetite whet, he decided to pursue a PhD from the Program in Molecular Biophysics at Johns Hopkins University. Following three training rotations, he scared away all potential mentors, except for Prof. Jeffrey J. Gray. In a shock to both parties, Jeliazko's child-like curiosity and energy paired excellently with Jeff's hands-off mentoring approach. Like the Tasmanian Devil, Jeliazko ran wild, pursuing the computational modeling of antibodies and a menagerie of interesting, to him, proteins that are reported in this thesis and in a number of publications. His next mentor is Prof. Andreas Plückthun of the University of Zürich Department of Biochemistry. Soon-to-be-Dr. Jeliazkov wishes his future advisor the best of luck.